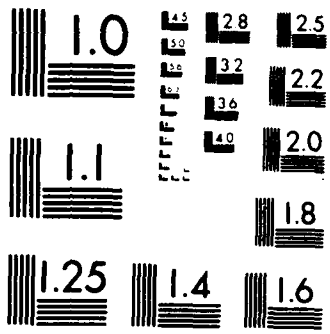
AD-A184 975 AIR FORCE OFFICER QUALIFYING TEST (AFOQT): ITEM AND FACTOR ANALYSIS OF FORM O(U) AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX J SKINNER ET AL AUG 87 1/1

UNCLASSIFIED AFHRL-TR-86-68 F/G 5/9 NL

END
10 87
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFHRL-TR-86-68

AIR FORCE

HUMAN RESOURCES

AD-A184 975

AIR FORCE OFFICER QUALIFYING TEST (AFOQT): ITEM AND FACTOR ANALYSIS OF FORM O

Jacobina Skinner
Malcolm James Ree

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

August 1987
Interim Technical Report for Period April 1986 - August 1986

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

This report has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

HAROLD G. JENSEN, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFHRL-TR-86-68	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Manpower and Personnel Division 6b. OFFICE SYMBOL (If applicable) AFHRL/MOAO	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (If applicable) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62703F PROJECT NO: 7719 TASK NO: 18 WORK UNIT ACCESSION NO: 47

11. TITLE (Include Security Classification)
Air Force Officer Qualifying Test (AFOQT): Item and Factor Analysis of Form O

12. PERSONAL AUTHOR(S)
Skinner, J.; and Ree, M.J.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM Apr 86 TO Aug 86	August 1987	58

16. SUPPLEMENTARY NOTATION
Analyses were conducted in AFHRL/TS Study number S9106.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	08	
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Air Force Officer Qualifying Test; item analysis; aptitude tests; item information curves; factor analysis; selection tests (Continued)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The item characteristics and factor structure of Form O of the Air Force Officer Qualifying Test (AFOQT) were investigated. Information on test difficulty, discrimination, reliability, speededness, and factor structure was obtained to guide developments and improvements in future versions of the AFOQT.

Test responses of a random sample of 3,000 officer applicants examined between September 1981 and September 1985 were analyzed. True score theory and Item Response Theory analytic techniques revealed that AFOQT Form O is a moderately difficult test with no extremely easy subtests. Most test items have adequate ability discrimination power. Items grouped in subtests are relatively homogeneous, and subtest scores have acceptable internal consistency reliability. Three of the 16 subtests in the battery fit the model of a power test, but most exhibit a speeded component. A factor analysis identified five ability dimensions which were judged to well represent the content of major aptitude composites derived from the AFOQT.

Recommendations to improve the structure of future forms included upgrading item discrimination and maximizing test information value by raising or lowering test difficulty in selected subtests to better match

4

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Office	22b. TELEPHONE (Include Area Code) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/TSR

Item 18 (Concluded):

test information curves

Item 19 (Concluded):

applicant ability. Parallel factor structure should be maintained in future tests. Follow-on research on samples of aircrew applicants is warranted.

SUMMARY

The Air Force Officer Qualifying Test (AFOQT) is a paper-and-pencil multiple aptitude battery used to select civilian applicants for officer precommissioning training programs and to classify commissionees into aircrew job specialties. A structural analysis of AFOQT Form O, the test version in operational use since September 1981, was conducted to support test development work on future versions. Information on item difficulty and discrimination, test reliability, speededness, and factor structure was obtained for guiding development and improvements in replacement tests and for assessing degree of continuity of test structure across forms.

Test responses were obtained from a random sample of 3,000 officer applicants administered AFOQT Form O between September 1981 and September 1985. Characteristics of test items were analyzed using true score theory and Item Response Theory analytic techniques. A factor analysis was conducted to evaluate the ability dimensions measured by the 16 subtests in the battery.

Results indicate that AFOQT Form O is a moderately difficult test. Item difficulties, subtest means, and negative skewness of raw score distributions reveal no extremely easy subtests. For most test items, item-test biserial correlations are high, indicating adequate ability discrimination. Subtests are composed of relatively homogeneous items, as reflected by subtest score reliabilities of .70 or higher. Three of the 16 subtests fit the model of a power test, but most exhibit a speeded component.

Five ability dimensions identified by factor analysis were labeled Verbal, Quantitative, Space Perception, Aircrew Interest/Aptitude, and Perceptual Speed. These factors were judged to closely approximate the content of major aptitude composites derived from the AFOQT.

Overall, AFOQT Form O appears to be a well-constructed test. Specific recommendations for improving future forms include upgrading item discrimination power. Test information value could be enhanced by adjusting the difficulty of selected subtests to better match applicant ability. In addition, replacement tests should be constructed to maintain the factor structure observed in AFOQT Form O. Finally, the current research stream should be continued with aircrew applicant samples tested on AFOQT Form O, as well as on both officer and aircrew applicant samples tested on future forms. Results would provide valuable information for assessing and improving the parallelism of test versions, the continuity of factor structure across forms, and the utility of the test for pilot and navigator selection.

PREFACE

The Air Force Human Resources Laboratory (AFHRL) is tasked as the test development agency for the Air Force Officer Qualifying Test (AFOQT) by Air Force Regulation 35-8, Air Force Military Personnel Testing System. The current research and development (R&D) effort was undertaken as part of AFHRL's responsibility to develop, revise, and conduct research in support of the AFOQT. Work was accomplished under Task 771918, Selection and Classification Technologies, which is part of a larger effort in Force Acquisition and Distribution Systems. The effort was completed under work unit 77191847, Development and Validation of Civilian and Nonrated Officer Selection Methodologies.

The authors acknowledge with gratitude the assistance of Mr. Jim Friemann and Ms. Suzanne Farrell of the AFHRL Technical Services Division. Their efforts were instrumental to the successful accomplishment of the data analysis phase of this effort. AlC Dave Lawson deserves special thanks for his outstanding contributions to data analysis and for being responsive to requests for changes to analyses.

TABLE OF CONTENTS

LIST OF FIGURES

List of Figures (Continued)

List of Figures (Concluded)

LIST OF TABLES

List of Tables (Concluded)

AIR FORCE OFFICER QUALIFYING TEST (AFOQT):
ITEM AND FACTOR ANALYSIS OF FORM O

I. INTRODUCTION

The Air Force Officer Qualifying Test (AFOQT) is a multiple aptitude battery with a developmental history dating to the early 1950s when the first version, Form A, was introduced. Since then, the test has been revised periodically to update items, to reduce test compromise opportunity, and to improve the predictive validity of the battery. The current operational version, Form O, is the sixteenth update in the AFOQT series and was implemented in September 1981.

AFOQT Form O, like its predecessors, is a paper-and-pencil instrument with multiple-choice test items designed for group administration under standardized conditions. As shown in Table 1, AFOQT Form O contains a total of 380 items distributed in sets of 15 to 40 items among 16 subtests. Each subtest is independently timed, with testing times varying between 3 and 29 minutes. Administration time for the entire battery is about 4.5 hours. Formally, all subtests are defined as power tests, although the completion rates by examinees in the standardization sample for the majority of the subtests suggest an underlying speeded component (Rogers, Roach, & Wegner, 1986). In general, the subtests are designed to assess verbal, quantitative, spatial, and specialized ability areas. A detailed description of the types of items in each subtest is presented in Appendix A.

Table 1. **Description of AFOQT Form O Subtests**

Subtest	Number of items	Testing time (minutes)	No. of items not reached by % applicants[a] 5%	20%
Verbal Analogies	25	8	6	2
Arithmetic Reasoning	25	29	9	4
Reading Comprehension	25	18	10	4
Data Interpretation	25	24	12	7
Word Knowledge	25	5	10	0
Math Knowledge	25	22	13	4
Mechanical Comprehension	20	22	0	0
Electrical Maze	20	10	15	10
Scale Reading	40	15	19	13
Instrument Comprehension	20	6	13	8
Block Counting	20	3	12	8
Table Reading	40	7	23	16
Aviation Information	20	8	5	0
Rotated Blocks	15	13	0	0
General Science	20	10	3	0
Hidden Figures	15	8	9	2

[a]Data are reproduced from Rogers et al. (1986) and are based on an AFOQT Form O equating/standardization sample of 37,409 cases.

Test score results are obtained by automated optical scanning of answer sheets and computerized scoring of item responses. Individual subtests are scored as "number right" by counting the number of items answered correctly. Subtest scores are then aggregated into

composite scores and reported on a percentile scale. The percentile scale reflects a normative group which was administered AFOQT Form N during the process of application for an Air Force commission during the late 1970s and early 1980s (Gould, 1978; Rogers et al., 1986). AFOQT Form O was then equated to Form N.

Five aptitude composites--Verbal, Quantitative, Academic Aptitude, Pilot, and Navigator-Technical--are derived. Table 2 shows the content of each and illustrates that the composites are constructed of partially overlapping sets of subtests. The Verbal and Quantitative composites each contain three subtests. These two composites are then combined to form the Academic Aptitude composite. The remaining 10 subtests are used exclusively in either the Pilot or Navigator-Technical composite or both. Additionally, the Navigator-Technical composite incorporates the Quantitative subtests, and the Pilot composite includes the Verbal Analogies subtest which is found also in the Verbal composite.

Table 2. Composition of AFOQT Form O Aptitude Composites

Subtest	Composite: Verbal	Quantitative	Academic Aptitude	Pilot	Navigator-Technical
Verbal Analogies	X		X	X	
Arithmetic Reasoning		X	X		X
Reading Comprehension	X		X		
Data Interpretation		X	X		X
Word Knowledge	X		X		
Math Knowledge		X	X		X
Mechanical Comprehension				X	X
Electrical Maze				X	X
Scale Reading				X	X
Instrument Comprehension				X	
Block Counting				X	X
Table Reading				X	X
Aviation Information				X	
Rotated Blocks					X
General Science					X
Hidden Figures					X

Throughout its history, the AFOQT has been used principally to select civilian applicants for officer precommissioning training and to classify commissionees into aircrew job specialties as pilots and navigators. Since its implementation in September 1981, the annual testing load on AFOQT Form O has been about 35,000 examinees, most of whom have been civilian applicants. The aptitude prerequisite for selection is set as a multiple cutoff on the Verbal and Quantitative composites. The standard applies to Officer Training School (OTS) and to Reserve Officer Training Corps (ROTC) commissioning programs; however, applicants for the third major commissioning program, at the Air Force Academy, are exempt from AFOQT requirements. Similar multiple-aptitude standards on the Pilot and Navigator-Technical composites are used as second-stage selectors for commissionees seeking aircrew job classifications. Secondary uses of the AFOQT include selection decisions for the Air Force Reserve (AFRES), Air National Guard (ANG), and the ROTC scholarship program.

The purpose of the present effort was a structural analysis of AFOQT Form O. Information at a sufficient level of detail to assure a full understanding of the test's structure has not been previously reported. Earlier publications on AFOQT Form O (Rogers et al., 1986), and on previous forms as well (e.g., Gould, 1978; Miller, 1974), have been limited primarily to a description of test development and standardization procedures. Test psychologists who work with the AFOQT on a daily basis and who are tasked with the construction of new forms need more in-depth knowledge of the AFOQT's characteristics at the item, subtest, and composite levels. Furthermore, greater familiarization is required to explicate fully the purpose and nature of the AFOQT, and its value to Air Force users and the military personnel testing community at large.

Presently, the Air Force Human Resources Laboratory (AFHRL) is engaged in the early stages of the test development cycle for AFOQT Form Q. Experimental versions of Form P, designed to be parallel to Form O, have already been completed and are scheduled for implementation in 1987. Structural analyses of Form O, together with results of planned similar analyses on AFOQT Form P data, will be used to assess the degree of continuity of test structure across forms. The present analyses will serve as a guide to changes and improvements in the item difficulty, discrimination, and factor structure of AFOQT Form Q.

II. METHOD

Subjects

Subjects were 3,000 examinees tested for operational purposes on AFOQT Form O. Records of AFOQT Form O responses on all examinees tested operationally are archived by the Technical Services Division (AFHRL/TS) to support research and development (R&D) efforts at AFHRL. When the present investigation began, records were available on 126,747 examinees with testing dates during the 4-year period between September 1981 and September 1985. Data editing checks resulted in the elimination of 14,049 records. The majority (11,310) were duplicate records of retesters from the second and/or subsequent administration of the AFOQT. An additional 2,684 records had incomplete or invalid data on background and demographic variables. Finally, 55 records without responses to any of the 380 test items were deleted. The analysis sample was drawn randomly from the remaining 112,698 records. The sample was limited to 3,000 subjects to accommodate the capacity of some of the analytic software and to facilitate processing.

Inspection of the sample characteristics, as shown in Table 3, revealed that the subjects were representative of the population of examinees testing for the first time on the AFOQT between 1981 and 1985. The sample was mixed by sex and race, with the majority of subjects being White males. Examinees were 22 years of age on the average (population mean = 22.28), and all had completed a secondary education program by degree or certificate of equivalency. About 39% held college degrees (baccalaureate or higher). The average number of years of education completed was 14.4 (population mean = 14.43). About 88% of the examinees were tested in conjunction with application to officer precommissioning training programs through OTS (44.5%) and ROTC (43.3%). The remainder were applying to ANG, AFRES, or other Air Force programs requiring records of AFOQT scores. The sample was diverse with respect to geographical location of test site. Examinees had tested at Military Entrance Processing Stations (MEPS), ROTC detachments located on college and university campuses, and at Consolidated Base Personnel Offices (CBPOs) on Air Force installations in the continental United States and overseas.

AFOQT Form O Scoring

Of the 380 items in the test booklet, a total of 12 items were deleted from operational scoring due to double keys, miskeys, or poor item performance. The same items were excluded from

the present analyses. By subtest, the number of omitted items was three in Verbal Analogies, four in Arithmetic Reasoning, two in Data Interpretation, one in Word Knowledge, one in Mechanical Comprehension, and one in Scale Reading. It should be noted that the number of items entered into analysis was reduced accordingly from the counts shown for these subtests in Table 1. Corresponding reductions occurred in the upper limit of subtest raw scores. For a full discussion of the development of AFOQT Form O, see Rogers et al. (1986).

Table 3. Percentage of AFOQT Form O Population (N = 112,698) and Sample (N = 3,000) by Demographic and Background Categories[a]

Sex	Population	Sample	Race	Population	Sample
Male	83.4	84.4	Black	12.6	12.4
Female	16.6	15.6	White	78.9	79.7
			Other	8.5	7.9
Degree	**Population**	**Sample**	**Program**	**Population**	**Sample**
High School	54.4	54.3	OTS	44.3	44.5
Associate's	7.6	7.0	ROTC	42.9	43.3
Bachelor's	36.0	36.9	ANG	4.1	4.0
Master's	1.8	1.8	Reserves	1.2	1.3
Doctoral	.0	.0	Other	7.5	6.9

[a]Column percentages may not sum to 100.0 due to rounding.

Analysis

Test structure was evaluated by analysis of both the items and the subtests. Two types of item analytic procedures were used. The first was based on the widely recognized classical or "true score" theory (Davis, 1951; Gulliksen, 1950; Henrysson, 1971). The second was an application of Item Response Theory (IRT), developed by Lord and Novick (1968). The subtests were evaluated using descriptive and correlational analyses, as well as factor analytic techniques.

True Score Item Analysis. Classical analyses were conducted to explore characteristics of items in each of the 16 subtests. Item difficulties (p) were calculated as the proportion of examinees responding correctly to the item. The biserial correlation (r_{bis}) between item score (correct or incorrect) and total test score (subtest raw score) was obtained as an index of the discrimination value of the item. Omitting was also investigated.

Item Response Theory (IRT) Analysis. Additional information on AFOQT items was obtained from IRT analyses. IRT describes items in terms of the likelihood of an item's being answered correctly at different examinee ability levels (often indicated by θ), and a set of estimates of parameters which describe a curve. This curve is referred to as an Item Characteristic Curve (ICC). It generally takes the shape of an ogive. In practice, the ogive can be described by three parameters and is based on the logistic approximation to the normal ogive. The logistic ogive is preferable due to its mathematical tractability.

Three parameters currently called a, b, and c are used to describe the curve. The item discrimination parameter (a) is a function of the slope of the ICC and generally ranges from .4 to about 2.5. A value of a equal to about 1.0 is typical of many test items; a values below .5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are

infrequently found. The item difficulty parameter (b) describes the point of inflection of the ICC and is usually scaled between -3.0 and +3.0, although the metric is arbitrary. The item guessing parameter (c) is the lower asymptote of the ICC and is generally interpreted as the probability of selecting the correct item-option by chance alone. Most test items have c parameters greater than 0.0 and less than or equal to .30.

Figure 1 shows three ICCs. The horizontal axis is scaled in units of ability (θ), and the vertical axis is the probability (P) of answering the item correctly ($P(1/\theta)$). The solid curved line shows an ICC for an item of average difficulty with acceptable discrimination and a lower asymptote appropriate for a five-option multiple-choice item. The dashed line shows an item of identical difficulty, with a c value of .28, but with a lower a value. Note how the slope of the curve is less steep. The third curve (dot-dash line) shows an item with a c value of .30, an a parameter of 1.0, and a b parameter equal to 1.0. As the b parameter changes, the location of the inflection point of the curve is displaced along the horizontal axis.

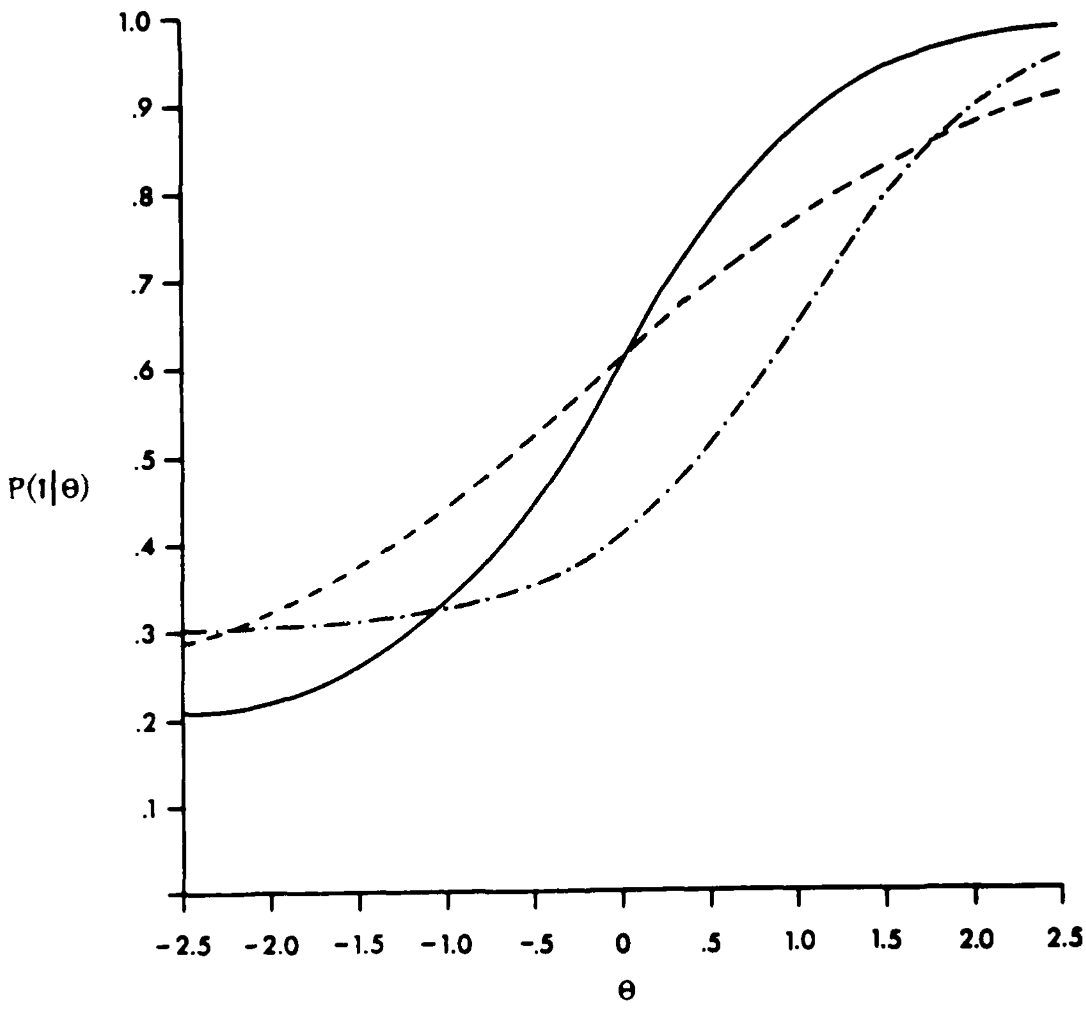


Figure 1. Item Characteristic Curves.

In most cases, the test constructor is faced with the task of estimating three parameters for each of the n items and one ability parameter (θ) for every examinee (N) such that N + 3n parameters must be estimated for each group of test items and examinees.

In the present effort, the computer program "LOGIST5" (Wingersky, Barton, & Lord, 1982), which computes joint maximum likelihood (JML) estimates of the item ($\underline{a}$, $\underline{b}$, $\underline{c}$) and ability (θ) parameters, was used. Previous research has evaluated the accuracy of parameter estimation by the JML method (Ree, 1979) and found it to perform adequately under proper conditions.

Subtest Analysis. Subtest raw scores were evaluated in terms of distributional shape, reliability, and structure. Estimates of the first four moments (mean, variance, skew, and kurtosis) were obtained to describe test score distributions. Test reliability or internal consistency was computed using Coefficient Alpha.

Factor analysis was used to assess the structure underlying the AFOQT subtests. Communality estimates were used in the principal diagonal of the matrix of subtest intercorrelations. A principal factors extraction was accomplished, as were both orthogonal and oblique rotations.

III. RESULTS AND DISCUSSION

Testing time limits are known to affect the measurement of abilities and are an important property of any aptitude test. The possible speeded nature of some AFOQT Form O subtests needed to be fully addressed, not only for test description purposes but also to guide critical analytic decisions, particularly those regarding the suitability of the subtests for IRT analysis. Knowledge of degree of speededness is also essential for adequate interpretation of results. This topic will be discussed in greater detail later in the report.

In Form M and Form N, the test versions immediately preceding Form O, each subtest was defined as a speeded or nonspeeded test (Gould, 1978; Miller, 1974). The majority of content areas in AFOQT Form O were carried forward from the earlier tests. Rogers et al. (1986) elected to treat all subtests as power (nonspeeded) tests, although many of the subtests were acknowledged to exhibit a speeded component. The speeded properties of five subtests--Instrument Comprehension, Scale Reading, Table Reading, Block Counting, and Electrical Maze--are of special interest in the present investigation, because Gould and Miller had specifically designated them as speeded tests.

Omitting rates were used to evaluate the power versus speed issue. Power tests typically have sufficient time limits to allow every examinee (in practice defined as 95%) to consider and answer every item. The proportion of the examinees who omitted items was plotted against item number (see Appendix B). A power test should show a low flat line with no more than about 5% of the examinees omitting on any item. The plot for a pure speeded test should begin low and straight and then increase toward the end of the test. When plotted, the subtests can then be categorized into three ideal types: power, speeded, and mixed.

Three of the subtests seemed to fit the model of a power test well: Mechanical Comprehension, Rotated Blocks, and General Science. The Electrical Maze, Instrument Comprehension, and Block Counting subtests appeared to be primarily speeded, whereas the other subtests were mixed (having more or less of a speeded factor).

There is nothing intrinsically wrong with using speeded or even mixed model subtests; however, the test constructor should specify the nature of the test, and it should turn out as specified. Whether those subtests which were previously specified as power or speed and which appeared clearly in the present analyses to better fit a mixed model have more or less predictive validity for officer selection is unknown. However, the speeded subtests on the Armed Services Vocational Aptitude Battery (ASVAB) (see Wegner & Ree, 1985) have been shown to be sensitive to timing, answer sheet type, and other environmental influences. Bearing this in mind, a conscious decision should be made, based on experience and empirical evidence, as to whether mixed model and/or speeded subtests should remain in the AFOQT.

The following commentary on speeded or mixed subtests is offered. In comparing the figures (curves) for omitting, it became evident that an index for speededness might be derived by fitting parameters to the curves. Consider trying to fit an ogive ($y = e^x$ form) to the three curves for the Mechanical Comprehension, Reading Comprehension, and Electrical Maze subtests. For Mechanical Comprehension, the fit will be relatively poor and the "slope" analog parameter will be almost zero. For Reading Comprehension, a slope parameter of moderate value will be found, but the inflection point would be estimated at a value greater than the number of items in the subtest. Finally, for Electrical Maze, both a high slope parameter and an inflection point within the range of possible subtest scores will be found. These two parameter estimates (inflection point and slope) might be used to characterize the amount of speededness in a test.

True Score Item Analysis

Information on the relative difficulty and discrimination of items in AFOQT Form 0 subtests was obtained using classical analysis procedures. Item difficulty (*p*) values are summarized in Table 4. Minimum and maximum item difficulties for each subtest are shown, followed by distributions of items in intervals from extremely difficult (≤ .20) to very easy (.81 to .99).

Table 4. Distribution of Item Difficulties (*p*) in AFOQT Form 0 Subtests

			Number[b] of items in range				
Subtest	Minimum	Maximum	≤ .20	.21 to .40	.41 to .60	.61 to .80	.81 to .99
Verbal Analogies	.29	.91	0	5	5	7	5
Arithmetic Reasoning	.26	.86	0	6	7	6	2
Reading Comprehension	.38	.81	0	1	9	14	1
Data Interpretation	.12[a]	.89	1	7	9	4	2
Word Knowledge	.24	.86	0	5	10	8	1
Math Knowledge	.39	.81	0	1	16	7	1
Mechanical Comprehension	.33	.83	0	2	12	4	1
Electrical Maze	.13[a]	.65	3	7	7	3	0
Scale Reading	.21	.92	0	13	14	9	3
Instrument Comprehension	.24	.65	0	7	11	2	0
Block Counting	.20	.91	1	7	4	5	3
Table Reading	.14[a]	.94	3	8	3	7	19
Aviation Information	.20	.73	1	7	10	2	0
Rotated Blocks	.27	.85	0	6	4	2	3
General Science	.09[a]	.78	1	7	10	2	0
Hidden Figures	.34	.93	0	3	5	2	5

[a]Value lower than would be expected from guessing on a power test.

[b]Items deleted in operational scoring are omitted from this table. See previous section titled AFOQT Form 0 Scoring.

Inspection of frequencies distributed in the five item difficulty intervals revealed that the bulk of the items in most subtests fell between .21 and .80. Most subtests were characterized by items of average (.41 to .60) or harder-than-average (.21 to .40) difficulty. The most striking finding was that nearly one-third of the subtests in the battery (5 of the 16) appeared to be quite difficult. Approximately 33 to 50% of the items in five subtests (Electrical Maze, Scale Reading, Block Counting, Rotated Blocks, and General Science) had difficulty values below .41. The Electrical Maze, Scale Reading, and Block Counting subtests showed large amounts of omitting,

which has tended to distort the estimated difficulty of items in these subtests; i.e., the items appear to be more difficult than they really are. All of the difficult items were found toward the end of each of these subtests. Rotated Blocks and General Science items did not show high levels of omitting and clearly appeared to be comprised of difficult subject matter. Only the Table Reading subtest contained a substantial number of extremely easy items; difficulty values for 19 of its 40 items were .81 or higher. All 19 of these easy items were at the beginning of the subtest. Almost all examinees attempted them and answered them correctly. Inspection of item content (see Appendix A) verified that they are indeed easy items. The item difficulties for the remaining 21 items were contaminated by speededness and probably do not reflect the true nature of the items.

Four subtests with items having difficulty values usually associated with guessing ($\leq .20$ for five-choice items)--that is, the most difficult items--were identified from inspection of minimum $\underline{p}$ values: Data Interpretation, Electrical Maze, Table Reading, and General Science. The General Science subtest contained the single most difficult item in the AFOQT battery ($\underline{p} = .09$). The Verbal Analogies, Block Counting, Table Reading, and Hidden Figures subtests each had at least one item in the very easy range ($\underline{p} > .91$).

Item discrimination data are presented in Table 5. In general, AFOQT test items appeared to discriminate quite well among examinees of relatively low and high ability levels. The distributions clearly showed that the majority of items in the subtests, with the notable exception of Scale Reading, had average (.41-.60) or above-average (.61-.80) discriminative power. No items performed extremely poorly, although the minimum values indicated that a few were operating at a marginal level; the Data Interpretation, Mechanical Comprehension, and Scale Reading subtests contained one or more items with discrimination values below .30. In the Scale Reading subtest, about 25% of the items (10 of 39) performed below average. The location of these items in the subtest did not appear to be a confounding factor. They are concentrated neither near the beginning nor end of the test, but are distributed throughout, suggesting that the items are simply weak discriminators. The finding of 17 items in the .81 to .99 interval for the Table Reading subtest merits discussion. The frequency was extremely high in relation to that for the other subtests (perhaps spuriously so) and may reflect the inflationary effects of the speededness on $\underline{r}_{bis}$ values (Henrysson, 1971).

Additionally, analyses were conducted to evaluate guessing. The proportion of the sample whose scores were below that expected by random guessing (often called the chance level score) was found by first dividing the number of items in each subtest by the number of item response choices. Items in all subtests except Instrument Comprehension have five response choices; Instrument Comprehension items have four. Then, examinees' scores were distributed, and the proportion at or below the chance level was computed. In a normative sample, the first percentile is usually comprised of examinees responding at the chance level. This cannot be expected in every other sample, although a large proportion of examinees scoring below the chance level usually indicates a mismatch between test difficulty and the examinees' ability level. Such a mismatch lowers the utility of the test in making decisions along the entire score continuum.

Table 6 shows the percentage of examinees who scored at or below chance level for all the subtests, as well as the ratio of the subtest means to chance scores. Values approaching 1.0 on this ratio indicate that the average response is near chance. The Instrument Comprehension subtest showed the highest proportion of scores at or below the chance level and was followed by the Electrical Maze subtest. It should be noted that these subtests are not used for general qualification for commissioning but rather, are used in more restricted samples for pilot and navigator selection. Similar analyses on applicant samples for aircrew training, which are typically composed of higher aptitude examinees, are needed to assess more fully the utility of these subtests.

Table 5. **Distribution of Item Discrimination Values (r_{bis}) in AFOQT Form 0 Subtests**

Subtest	Minimum	Maximum[a]	Number[b] of items in range .21 to .40	.41 to .60	.61 to .80	.81 to .99
Verbal Analogies	.43	.78	0	13	9	0
Arithmetic Reasoning	.45	.69	0	10	11	0
Reading Comprehension	.47	.83	0	7	15	3
Data Interpretation	.29	.66	6	16	1	0
Word Knowledge	.45	.80	0	5	19	0
Math Knowledge	.43	.84	0	9	15	1
Mechanical Comprehension	.24	.68	4	10	5	0
Electrical Maze	.36	.85	1	7	11	1
Scale Reading	.27	.64	10	26	3	0
Instrument Comprehension	.47	.81	0	8	11	1
Block Counting	.45	.79	0	6	14	0
Table Reading	.41	.88	0	11	12	17
Aviation Information	.38	.70	2	14	4	0
Rotated Blocks	.54	.76	0	3	12	0
General Science	.30	.72	5	10	5	0
Hidden Figures	.49	.70	0	8	7	0

[a]Values above .65 are quite high for a power test.

[b]Items deleted in operational scoring are omitted from this table. See previous section titled AFOQT Form 0 Scoring.

Table 6. **Percent of Examinees Scoring at or Below Chance Level and Ratio of Mean Subtest Scores to Chance Level**

Subtest	Percent at or below chance level	Subtest mean/ chance level
Verbal Analogies	2.87	3.34
Arithmetic Reasoning	6.97	2.75
Reading Comprehension	4.57	3.17
Data Interpretation	3.23	2.79
Word Knowledge	9.00	2.66
Math Knowledge	7.00	2.90
Mechanical Comprehension	7.43	2.45
Electrical Maze	23.70	1.92
Scale Reading	4.17	2.51
Instrument Comprehension	28.70	1.76
Block Counting	8.97	2.66
Table Reading	1.83	3.31
Aviation Information	13.47	2.16
Rotated Blocks	12.97	2.53
General Science	11.90	2.14
Hidden Figures	2.10	3.20

IRT Analysis

The existence of speeded and mixed model subtests presents a problem vis-à-vis planned IRT analyses. Strictly speaking, only the power subtests are amenable to IRT analyses. Adding the

speed factor causes a violation of the assumption of unidimensionality of extant IRT models. The consequences of using unidimensional models on multidimensional data are difficult to specify; in fact, detecting multidimensionality is an uncertain process. Nonetheless, IRT analyses were executed on all the subtests. Only for the Electrical Maze and Block Counting subtests, both of which were identified earlier as pure speeded tests, did the estimates fail to converge. For all other subtests including Table Reading, a third speeded test, and those designated as mixed or pure power tests, the estimates converged. IRT analyses of Mechanical Comprehension, Rotated Blocks, and General Science (clearly shown to be power subtests) were generally appropriate. Caution should be exercised in interpreting the IRT analyses of all but the power subtests.

The ICC parameters were computed using the JML procedure implemented in LOGIST5 (Wingersky et al., 1982). In all but two cases (Electrical Maze and Block Counting), the program reached convergence although certain item parameters--notably the $\underline{c}$--frequently showed default (median of estimable $\underline{c}$ parameters) values. Table 7 provides descriptive statistics for each of the parameters as estimated for each of the subtests on which estimates could be derived. Table 8 shows the distribution of estimated $\underline{a}$ parameters by subtest, and Tables 9 and 10 show the distribution of estimated $\underline{b}$ and $\underline{c}$ parameters.

It may be observed in Tables 7 through 10 that the distributions of estimated item parameters were all as expected. There were no surprises and no truly extreme values. In six of the 12 subtests, the $\underline{a}$ parameter went to the specified maximum default value (+2.0) one or two times, however (which is fewer than experience would predict). Also, the minimum $\underline{a}$ values were lower ($\leq$ .30) than usually expected, especially so in Mechanical Comprehension and somewhat so in Aviation Information, Data Interpretation, Scale Reading, and General Science. The very low $\underline{a}$ estimates were associated with very easy items, which are notoriously difficult to estimate. The standard errors of the estimated parameters were generally low.

In general, the $\underline{b}$ parameters were estimated well, with low estimated standard errors. Eight of the 12 subtests had minimum-maximum values between approximately $\pm$2.50 theta units, a range easily estimated. Mechanical Comprehension showed the most extreme values (-3.25 to +3.05), while Math Knowledge had the smallest range and was followed closely by Reading Comprehension. The Hidden Figures and Mechanical Comprehension subtests showed the greatest variability of $\underline{b}$. The single highest value of $\underline{b}$ was found in Data Interpretation and the lowest, in Table Reading.

For most of the subtests, approximately one-third of the $\underline{c}$ parameters could not be estimated and were assigned the default value of the median of the estimable $\underline{c}$ parameters in the subtest (Wingersky et al., 1982). It is interesting to note that Data Interpretation, which had the greatest range of estimated $\underline{c}$ parameters, also had a large range of estimated $\underline{b}$ parameters. This was most likely the result of the estimation process implemented in LOGIST5 which yields correlated estimates of the item parameters. The results of the IRT analyses were generally consistent with the results of the classical item analyses. Difficult items had the higher $\underline{b}$ values, and easy items showed low $\underline{b}$ values.

Correlations of the proportion of examinees omitting an item and the IRT item parameters were computed and are presented in Table 11. For five of the 14 subtests, the highest correlation was found for the $\underline{a}$ parameter. Mechanical Comprehension, the subtest which best fits the model of a power test, showed the lowest level of correlations, with omitting most closely related to difficulty. This is consistent with theory. Table Reading, a speeded test, was conspicuous in its generally high level of correlation between omitting and the IRT item parameters.

Table 7. Descriptive Statistics of IRT Parameters for AFOQT Form O Subtests

Subtest	Item discrimination (a)				Item difficulty (b)				Item guessing (c)			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
Verbal Analogies	.93	.25	.47	1.45	-.24	1.18	-2.13	1.35	.14	.07	.03	.33
Arithmetic Reasoning	1.10	.33	.40	2.00	.19	1.00	-1.52	1.69	.17	.07	.09	.36
Reading Comprehension	1.11	.55	.43	2.00	-.36	.73	-1.75	.90	.14	.07	.03	.29
Data Interpretation	.83	.43	.27	1.92	.40	1.45	-3.11	3.23	.17	.10	.00	.47
Word Knowledge	1.17	.43	.41	1.89	-.01	.96	-2.83	1.89	.15	.06	.03	.27
Math Knowledge	1.13	.48	.36	2.00	-.06	.72	-1.56	1.08	.16	.11	.02	.42
Mechanical Comprehension	.81	.39	.11	1.49	.34	1.29	-3.25	3.05	.15	.06	.06	.29
Scale Reading	.92	.56	.25	2.00	.08	1.26	-3.42	1.43	.17	.08	.09	.38
Instrument Comprehension	1.25	.53	.41	2.00	.45	.83	-1.49	2.10	.20	.13	.03	.43
Table Reading	1.39	.64	.35	2.00	-.98	1.78	-3.83	2.26	.14	.06	.08	.26
Aviation Information	.88	.37	.29	1.66	.51	.96	-2.08	2.02	.13	.04	.07	.23
Rotated Blocks	1.10	.34	.57	1.74	.14	1.05	-1.88	1.33	.10	.09	.00	.31
General Science	.94	.42	.21	1.87	.74	.90	-1.10	2.51	.15	.09	.05	.38
Hidden Figures	.93	.34	.57	1.61	-.51	1.41	-2.48	1.29	.18	.07	.11	.31

Table 8. Distribution of IRT Item Discrimination Parameter (a) for AFOQT Form O Subtests

Subtest	Number of items in range				
	<.40	.40 to .80	.81 to 1.00	1.01 to 1.40	>1.40
Verbal Analogies	0	5	7	9	1
Arithmetic Reasoning	0	2	6	9	4
Reading Comprehension	0	10	4	3	8
Data Interpretation	5	7	2	7	2
Word Knowledge	0	7	5	4	8
Math Knowledge	1	5	7	5	7
Mechanical Comprehension	3	7	2	5	2
Scale Reading	6	16	6	3	8
Instrument Comprehension	0	5	3	4	8
Table Reading	1	12	0	6	21
Aviation Information	1	7	7	2	3
Rotated Blocks	0	3	2	8	2
General Science	2	6	4	6	2
Hidden Figures	0	6	5	2	2

Insofar as speededness represents a second trait, it violates the unidimensionality assumption of IRT and causes parameter estimates to be biased or degraded in some fashion. Empirical ICCs were plotted for several of the subtests, and all were representative of appropriate unidimensional ICCs. This is theoretically troublesome and requires more study and evaluation.

Test Information Curves (TICs)

TICs were computed for each subtest from its estimated ICC parameters for the range of $-3.0 \leq \theta \leq +3.0$. These curves show the amount of Fisher information at each interval of θ. The vertical axis (Information) is the first partial derivative divided by the conditional variance of θ. For these estimates, θ was assumed to be known; and a, b, and c were treated as true values.

Appendix C shows the TIC for each subtest. The height of the curve at each value of θ is often interpreted as a conditional measure of reliability; that is, the greater the height of the ordinate, the greater the reliability at that point.

The remarkable feature found for these tests was the relative flatness of most of the TICs and their usually unimodal shape. The curve for Math Knowledge looked like a textbook example of a peaked linear test; the curves for Word Knowledge, Reading Comprehension, Instrument Comprehension, and Rotated Blocks were similar but not quite so centered on the ability scale. The curve for Verbal Analogies reached its maximum below mean (and median) θ; therefore, it may be that the test is not providing maximum information for the average officer applicant. The opposite was true for Arithmetic Reasoning, Data Interpretation, Mechanical Comprehension, Instrument Comprehension, Rotated Blocks, Aviation Information, General Science, and Hidden Figures. Their TICs reached their maxima above the median of θ. The TIC for Hidden Figures showed a very broad spread and low information function for applicants. Its test information value for navigator selection should be investigated further. It might also be that the nonunidimensional nature (i.e., speededness) of the Hidden Figures subtest caused it to fail to meet the assumptions of the IRT model.

Table 9. Distribution of IRT Item Difficulty Parameter (b) for AFOQT Form O Subtests

Subtest	Number of items in range											
	< -2.50	-2.50 to -2.00	-1.99 to -1.50	-1.49 to -1.00	-.99 to -.50	-.49 to .00	.01 to .50	.51 to 1.00	1.01 to 1.50	1.51 to 1.90	1.91 to 2.50	> 2.50
Verbal Analogies	0	2	2	2	5	1	2	4	4	0	0	0
Arithmetic Reasoning	0	0	1	3	1	2	6	3	3	2	0	0
Reading Comprehension	0	0	1	6	3	6	6	3	0	0	0	0
Data Interpretation	1	2	0	0	0	5	3	5	4	0	2	1
Word Knowledge	1	0	0	2	5	2	4	8	1	1	0	0
Math Knowledge	0	0	1	3	2	4	11	3	1	0	0	0
Mechanical Comprehension	1	0	0	0	1	5	3	4	3	0	1	1
Scale Reading	2	1	2	3	3	5	3	8	12	0	0	0
Instrument Comprehension	0	0	0	2	0	2	5	8	2	0	1	0
Table Reading	12	2	3	4	2	4	3	3	2	3	2	0
Aviation Information	0	1	0	1	0	3	5	5	2	2	1	0
Rotated Blocks	0	0	3	0	0	1	3	5	3	0	0	0
General Science	0	0	0	1	1	1	5	5	2	4	0	1
Hidden Figures	0	3	2	1	1	1	1	3	3	0	0	0

Table 10. Distribution of IRT Item Guessing Parameter (c) for AFOQT Form O Subtests

Subtest	Number of items in range .00 to .10	.11 to .20	.21 to .30	>.30
Verbal Analogies	5	12	4	1
Arithmetic Reasoning	6	11	3	1
Reading Comprehension	14	5	6	0
Data Interpretation	3	15	4	1
Word Knowledge	11	9	4	0
Math Knowledge	11	6	5	3
Mechanical Comprehension	4	11	4	0
Scale Reading	1	27	9	2
Instrument Comprehension	6	4	3	7
Table Reading	22	10	8	0
Aviation Information	4	15	1	0
Rotated Blocks	9	4	1	1
General Science	6	10	2	2
Hidden Figures	0	11	3	1

Table 11. Correlation of Omits with IRT a, b, and c Parameters

Subtest	a	b	c
Verbal Analogies	.171	.535	.669
Arithmetic Reasoning	.457	.819	.549
Reading Comprehension	.778	.766	.871
Data Interpretation	.462	.625	.756
Word Knowledge	.490	.232	.345
Math Knowledge	.476	.582	.776
Mechanical Comprehension	-.063	.424	.047
Scale Reading	.916	.661	.782
Instrument Comprehension	.806	.734	.696
Table Reading	.803	.952	.813
Aviation Information	.336	.234	.140
Rotated Blocks	.371	.261	.209
General Science	.373	.425	.205
Hidden Figures	.698	.900	.729

Ability Distributions and Test Information

Using the maximum likelihood estimates of ability generated in the IRT analyses, polygons of the distribution were plotted. These are presented in Appendix D. For the most part, the polygons were quite unremarkable. One interesting phenomenon was that in the measures of general ability (Verbal Analogies, Arithmetic Reasoning, Reading Comprehension, Word Knowledge, and Math Knowledge), there were peaks toward the upper end of the distribution as if they might represent special subsamples. These peaks were not displayed by the special knowledge tests but did occur in two perceptual tests--Rotated Blocks and Hidden Figures. The nature of these subsamples awaits further analyses.

If the Test Information Curves and the distribution of ability polygons for a test are superimposed, information about the appropriateness of the test for a group of examinees is revealed. An appropriate test will have a distribution of ability in the same general shape as the TIC. If the modes of the two distributions, for example, do not coincide, then the test information is not appropriate for the examinees.

Verbal Abilities showed an acceptable fit of test to examinees, as did Table Reading, Math Knowledge, Rotated Blocks, Reading Comprehension, and Mechanical Comprehension (in decreasing order of goodness). The remaining subtests showed poor coordination between information and examinee ability, with Scale Reading showing the greatest discrepancy and the subtests used for pilot and navigator screening showing a generally poor relationship between test information and examinee ability. This may not be a problem because the subtests are used only for those expressing interest in pilot or navigator training. Further analyses with samples of applicants for pilot and navigator training are required to illuminate the issue.

Surprisingly, Word Knowledge showed a relatively poor match. This can be corrected in future forms through selection of appropriate items. Most likely, this can be accomplished by reducing the difficulty and providing easier items in future forms of the Word Knowledge subtest.

Subtest Analyses

Table 12 presents the results of the descriptive analyses of the scores of the AFOQT subtests. Shown in the table are the number of items scored, and the mean, standard deviation, skewness, kurtosis, and internal consistency estimate of reliability.

Table 12. **Descriptive Statistics of AFOQT Form O Subtest Scores**

Subtest	Number of items	Mean	SD	Skew	Kurtosis	Reliability[a]
Verbal Analogies	22	13.36	4.23	-.39	-.40	.80
Arithmetic Reasoning	21	11.00	4.40	.07	-.[illegible]5	.81
Reading Comprehension	25	15.83	5.93	-.30	-.93	.88
Data Interpretation	23	11.15	3.93	.18	-.36	.71
Word Knowledge	24	13.28	5.83	.08	-.99	.88
Math Knowledge	25	14.48	6.04	-.04	-1.07	.88
Mechanical Comprehension	19	9.78	3.65	.01	-.58	.71
Electrical Maze	20	7.68	4.22	.75	.24	.81
Scale Reading	39	20.07	6.73	-.03	-.37	.84
Instrument Comprehension	20	8.82	4.76	.36	-.69	.84
Block Counting	20	10.62	4.39	-.08	-.58	.83
Table Reading	40	26.46	7.35	-.50	.50	.92
Aviation Information	20	8.65	4.08	.56	-.16	.77
Rotated Blocks	15	7.59	3.36	-.06	-.76	.77
General Science	20	8.54	3.66	.42	-.29	.70
Hidden Figures	15	9.60	2.76	-.32	.03	.69

[a]Reliability estimated using Coefficient Alpha.

Findings with respect to subtest mean scores paralleled previously presented results on item difficulty (_p_ and _b_). They revealed a moderately difficult multiple aptitude battery with several rather difficult subtests (e.g., General Science and Electrical Maze) and no very easy subtests. The standard deviations generally increased or decreased with the number of items, as would be expected.

Most of the subtests showed a moderate degree of skewness, with the Electrical Maze subtest showing the greatest skewness (.75). Although some of the other subtests showed departures from normal kurtosis, none seemed particularly extreme. Math Knowledge was the most kurtotic of the subtests, with a flatter-than-normal distribution.

Reliabilities of the AFOQT subtests were all acceptable. The present results were comparable to those reported for the AFOQT Form O equating/standardization sample (Rogers et al., 1986). In both samples, Coefficient Alpha values computed on mixed model and speeded subtests should be considered inflated estimates of the true reliability. A technically more appropriate method of reliability estimation would involve the use of correlation between separately timed parallel forms. Such data are not available.

Subtest Intercorrelations and Factor Analysis

The matrix of intercorrelations of subtests presented in Table 13 shows a set of positively intercorrelated subtests. The highest correlation obtained (.77) was that between the Reading Comprehension and Word Knowledge subtests, both of which assess verbal aptitude. The lowest correlation (.17) was between the Word Knowledge subtest and Electrical Maze, a spatial-perceptual subtest. The Electrical Maze subtest, possibly due to its rather unique content, showed only low to moderate correlations with any of the other subtests; it showed the highest intercorrelation with Block Counting, another spatial-perceptual subtest. In general, the verbal subtests showed higher intercorrelations with other verbal subtests than with the nonverbal subtests. The same trend was observed for the quantitative subtests and for the spatial-perceptual subtests. These findings and the results on omitting suggested that at least three factors could be expected to emerge from a factor analysis--verbal, quantitative, and spatial--and possibly a fourth which taps speededness.

A principal factors analysis was conducted. The communalities were the squared multiple correlation of each subtest as predicted by all the other subtests. After inspection of solutions involving one through six factors, five factors were judged to best represent the data and were extracted and rotated both orthogonally by the Varimax method and by the Kaiser-Harris Type 2 oblique method. The oblique rotation method was more interpretable and was accepted as the appropriate solution.

Table 14 shows the obliquely rotated factor loadings and the matrix of intercorrelations among the factors. The intercorrelations of the factors were somewhat lower than expected with this type of oblique rotation. This suggested that a higher-order factor analysis would lead to the extraction of two or three higher-order factors rather than the one or two expected for most multiple aptitude batteries.

Table 15 shows the rankings of subtest factor loadings after the oblique rotation. The first and second factors are clearly Verbal and Quantitative, respectively; in fact, they closely approximate the Verbal and Quantitative composites of the battery. Factor three is composed of the spatial subtests and the Mechanical Comprehension subtest (which also depends on spatial ability to some extent) and has been termed the Space Perception factor. Factor four has four subtests which load above +.30, with the Aviation Information and Instrument Comprehension subtests best defining the factor, and is identified as the Aircrew Interest/Aptitude factor. Factor five is a second perceptual factor, which includes some of the most speeded of the subtests, and is identified as Perceptual Speed.

Table 13. Intercorrelation Matrix of AFOQT Form O Subtests

	VA	AR	RC	DI	WK	MK	MC	EM	SR	IC	BC	TR	AI	RB	GS	HF
Verbal Analogies (VA)	1.00															
Arithmetic Reasoning (AR)	.58	1.00														
Reading Comprehension (RC)	.73	.58	1.00													
Data Interpretation (DI)	.53	.67	.55	1.00												
Word Knowledge (WK)	.68	.46	.77	.46	1.00											
Math Knowledge (MK)	.55	.71	.51	.60	.40	1.00										
Mechanical Comprehension (MC)	.48	.51	.46	.46	.40	.48	1.00									
Electrical Maze (EM)	.27	.37	.23	.38	.17	.40	.44	1.00								
Scale Reading (SR)	.48	.66	.45	.62	.37	.60	.48	.45	1.00							
Instrument Comprehension (IC)	.34	.41	.33	.43	.28	.39	.49	.44	.49	1.00						
Block Counting (BC)	.45	.53	.40	.51	.32	.49	.50	.47	.61	.49	1.00					
Table Reading (TR)	.34	.44	.35	.47	.27	.44	.30	.31	.56	.34	.51	1.00				
Aviation Information (AI)	.30	.31	.34	.34	.32	.25	.50	.29	.33	.56	.31	.21	1.00			
Rotated Blocks (RB)	.43	.47	.35	.42	.29	.49	.54	.42	.49	.46	.55	.34	.34	1.00		
General Science (GS)	.51	.49	.55	.44	.51	.52	.57	.34	.41	.41	.37	.25	.46	.40	1.00	
Hidden Figures (HF)	.40	.40	.36	.39	.31	.40	.39	.34	.47	.36	.45	.36	.27	.42	.34	1.00

Table 14. Obliquely Rotated Factor Loadings and Intercorrelation of Factors

Subtest	Factor I	II	III	IV	V
Verbal Analogies	.626	.173	.227	-.012	.028
Arithmetic Reasoning	.177	.598	.058	.053	.176
Reading Comprehension	.740	.147	.028	.071	.095
Data Interpretation	.201	.401	-.003	.122	.332
Word Knowledge	.774	.013	.027	.091	.055
Math Knowledge	.115	.622	.172	.011	-.057
Mechanical Comprehension	.130	.192	.356	.393	-.074
Electrical Maze	-.119	.154	.364	.204	.124
Scale Reading	.046	.318	.188	.071	.439
Instrument Comprehension	-.017	.020	.172	.512	.258
Block Counting	.053	.084	.459	.045	.334
Table Reading	.068	.129	.155	-.045	.479
Aviation Information	.110	-.041	.016	.646	.099
Rotated Blocks	.027	.159	.510	.142	.031
General Science	.303	.291	.136	.381	-.153
Hidden Figures	.135	.050	.364	.044	.192

Intercorrelation of Factors

I	1.00				
II	.41	1.00			
III	.26	.49	1.00		
IV	.26	.30	.45	1.00	
V	.22	.48	.50	.24	1.00

Table 15. Ranks of Loadings[a] for the Rotated Factor Matrix

Subtest	Factor I	II	III	IV	V
Verbal Analogies	3				
Arithmetic Reasoning		2			
Reading Comprehension	2				
Data Interpretation		3			4
Word Knowledge	1				
Math Knowledge		1			
Mechanical Comprehension			5	3	
Electrical Maze			4		
Scale Reading		4			2
Instrument Comprehension				2	
Block Counting			2		3
Table Reading					1
Aviation Information				1	
Rotated Blocks			1		
General Science	4			4	
Hidden Figures			3		

[a]No subtest with loading less than |.30| ranked.

As with other multiple aptitude batteries, both Verbal and Quantitative factors were found. Additionally, two spatial factors frequently found in other multiple aptitude batteries were found. The other factor, Aircrew Interest/Aptitude, seems to be unique to the AFOQT. To a large degree, the factors replicate the composites used in operational commissioning systems.

IV. CONCLUSIONS

Several major conclusions were supported by the present analyses of the characteristics of AFOQT Form O. Results substantiated and added to the findings reported by Rogers et al. (1986) on the equating/standardization sample for AFOQT Form O. The knowledge base about the test was expanded, and important new insights were gained about item characteristics and test structure. The present findings have implications for future test forms and point to the need for additional research to provide definitive guidelines for test modifications.

The qualification test used to assess applicants' potential for success in Air Force officer precommissioning and aircrew training programs was found to contain a challenging mix of content areas, some of which are common to most multiple aptitude batteries and others which assess relatively unique abilities. For the most part, the test is moderately difficult. None of the subtests was found to be exceptionally easy for the sample of officer applicants. A few subtests such as General Science, Aviation Information, and Electrical Maze are quite difficult, probably due to their highly specialized or unique subject matter.

In general, AFOQT test items discriminate adequately among applicants of differing ability levels, and the subtests have acceptable internal consistency reliability. Scale Reading was the only subtest having an excessive number of items on which performance did not relate well to total test score. These items are not suitable for use as anchor items in future forms. Instead, new items with greater discriminative power should be substituted. It should be noted that the use of Coefficient Alpha as an indicator of internal consistency reliability probably produced overestimates of true reliability on the speeded and mixed model subtests. The construction of Form P in two versions will provide the opportunity to obtain better estimates of AFOQT subtest reliabilities with alternate forms.

The IRT analyses suggest that the mismatch between test information and the ability of examinees is sufficient to indicate the potential to improve the utility of the test. Reduction of mismatches can be expected to lead to higher validity of selection decisions. The six subtests that comprise the Verbal and Quantitative composites on which precommissioning entry standards are set are of primary concern. In three of these subtests (Verbal Analogies, Math Knowledge, and Reading Comprehension), test information matches the ability of officer applicants well; but the match could be improved in the other three subtests (Word Knowledge, Data Interpretation, and Arithmetic Reasoning), by making them less difficult through the use of easier items. However, such adjustments based only on the present analyses might be premature. The Verbal and Quantitative composites contain subtests found also in the Pilot and Navigator-Technical composites (see Table 2). Thus, test construction decisions on these subtests, and on the 10 additional subtests used exclusively for aircrew selection, should consider the fit between test information and examinee ability for samples of applicants for pilot and navigator training. Research using the appropriate samples is recommended as part of follow-on forms development activities.

Factor analytic results suggest that AFOQT Form O measures several ability dimensions. Verbal and Quantitative factors were clearly isolated. These results are encouraging for two reasons. First, and most important, they lend support and credibility to the like-named major aptitude composites which are derived from the aptitude battery for use as officer precommissioning entry

standards. Second, the presence of Verbal and Quantitative factors together suggests adequate test coverage of the general ability factor, G, the most universally predictive measure of ability (U. S. Employment Service, 1983).

The test content also encompasses additional major ability dimensions, as indicated by the three factors labeled Space Perception, Aircrew Interest/Aptitude, and Perceptual Speed. The Pilot and Navigator-Technical composites used for aircrew selection decisions align well with these empirically derived ability factors. Subtests in the Pilot composite overlap to a large degree with factors four and five (Aircrew Interest/Aptitude, and Perceptual Speed). The content of the Navigator-Technical composite appears to be best defined by factors two, three, and five (Quantitative, Space Perception, and Perceptual Speed). Current findings on factorial structure add to evidence of the construct validity of the aircrew composites. Recent studies have demonstrated the predictive validity of the Pilot composite for Undergraduate Pilot Training (Bordelon & Kantor, 1986) and of the Navigator-Technical composite for Undergraduate Navigator Training (Shanahan & Kantor, 1986).

Findings on response omitting rates have implications for test description and validity issues. Subtests retained in subsequent AFOQT forms should be labeled consistently and accurately as power, speed, or mixed model subtests. In Forms M and N, the Instrument Comprehension, Scale Reading, Table Reading, Block Counting, and Electrical Maze subtests appear to have been appropriately labeled as speeded tests, but current results suggest that the speeded label applies only to the latter three subtests while Instrument Comprehension and Scale Reading better fit the mixed model in Form O. It is recommended that the 95% completion rule be adopted as the operational definition of power subtests in the AFOQT. Subtests which fail to meet this definition would be most accurately described as mixed or speeded. As part of the test development cycle for follow-on forms, research is recommended on the subtest configuration (power, speed, mixed) which optimizes reliability and criterion-related validity. The relative contributions of speed, power, and a combination of speed and power to the utility of each subtest should be evaluated. These data would aid in determining whether mixed model, pure speed, or pure power subtests should be retained in the AFOQT.

Test developers should be cognizant of the sensitivity of test results on speeded and mixed model AFOQT subtests and should be alert to opportunities to mitigate contaminating influences. A major change planned for Forms P serves as a pertinent example. All examinees will use the same type of answer sheet. The new procedure will control for systematic error which may have been introduced by the use of different answer sheets for examinees tested on Form O at ROI and OTS test sites (Rogers et al., 1986). Other remedial activities would include educating test administrators and proctors about the particular importance of standardized test conditions for speeded and mixed model subtests to control timing and environmental contaminants.

A critical first step in the developmental cycle for follow-on forms of the AFOQT has been completed. The present analyses provide valuable information on the item characteristics and factor structure of AFOQT Form O for officer applicants. As previously mentioned, supplementary data are needed for both pilot and navigator applicants. Evaluation of the extent to which the test structure holds for other major groups, especially ethnicity and gender groupings, would be helpful. Continuation of the same research stream, as planned for AFOQT Forms P samples, is an essential activity. Collectively, results should provide a definitive basis for specifying the desired test characteristics of the AFOQT. Expected benefits include ensuring that test utility for officer and aircrew applicants is optimized, that equivalence of test versions is achieved, and that continuity of factor structure across forms is maintained.

REFERENCES

Bordelon, V.P., & Kantor, J.E. (1986). Utilization of psychomotor screening for USAF pilot candidates: Independent and integrated selection methodologies (AFHRL-TR-86-4, AD-A170 353). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Davis, F.B. (1951). Item selection techniques. In E. L. Lindquist (Ed.), Educational Measurement (pp. 266-328). Washington, DC: American Council on Education.

Gould, R.B. (1978). Air Force Officer Qualifying Test Form N: Development and standardization (AFHRL-TR-78-43, AD-A059 746). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley & Sons, Inc.

Henrysson, S. (1971). Gathering, analyzing and using data on test items. In R. L. Thorndike (Ed.), Educational Measurements (pp. 130-159). Washington, DC: American Council on Education.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental scores. Reading, MA: Addison-Wesley Publishing Co.

Miller, R.E. (1974). Development and standardization of the Air Force Officer Qualifying Test Form M (AFHRL-TR-74-16, AD-A778 837). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Ree, M.J. (1979). Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-385.

Rogers, D.L., Roach, B.W., & Wegner, T.G. (1986). Air Force Officer Qualifying Test Form O: Development and standardization (AFHRL-TR-86-24, AD-A172 037). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Shanahan, F.M., & Kantor, J.E. (1986). Basic Navigator Battery: An experimental selection composite for undergraduate navigator training (AFHRL-TR-86-3, AD-A168 857). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

United States Employment Service (1983). Overview of validity generalization for the U. S. Employment Service (USES Test Research Report No. 43). Washington, DC: Department of Labor.

Wegner, T.G., & Ree, M.J. (1985). Armed Services Vocational Aptitude Battery: Correcting the speeded subtests for the 1980 youth population (AFHRL-TR-85-14, AD-A158 823). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Winkersky, M., Barton, M., & Lord, F. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

APPENDIX A: DESCRIPTION OF AFOQT SUBTESTS

Table A-1. Description of Items in AFOQT Form O Subtests

Subtest	No. of items	Description
Verbal Analogies	25	Measures ability to reason and recognize relationships between words.
Arithmetic Reasoning	25	Measures ability to understand and reason with arithmetic relationships through word problems.
Reading Comprehension	25	Measures ability to read and understand paragraphs.
Data Interpretation	25	Measures ability to interpret data from graphs and charts.
Word Knowledge	25	Measures ability to understand written language through use of synonyms.
Math Knowledge	25	Measures ability to use learned mathematical terms, formulas, and relationships.
Mechanical Comprehension	20	Measures mechanical knowledge and understanding of mechanical functions.
Electrical Maze	20	Measures spatial ability to choose a correct path through a maze.
Scale Reading	40	Measures ability to read scales and dials.
Instrument Comprehension	20	Measures ability to determine aircraft attitude from flight instruments.
Block Counting	20	Measures spatial ability to "see into" a three-dimensional pile of blocks.
Table Reading	40	Measures ability to read tables quickly and accurately.
Aviation Information	20	Measures knowledge of general aeronautical concepts and terminology.
Rotated Blocks	15	Measures spatial aptitude by visualizing and manipulating objects in space.
General Science	20	Measures knowledge and understanding of scientific terms, concepts, principles, and instruments.
Hidden Figures	15	Measures perceptual and visual imagery ability using simple figures embedded in complex drawings.

APPENDIX B: OMITTING CURVES FOR THE SUBTESTS

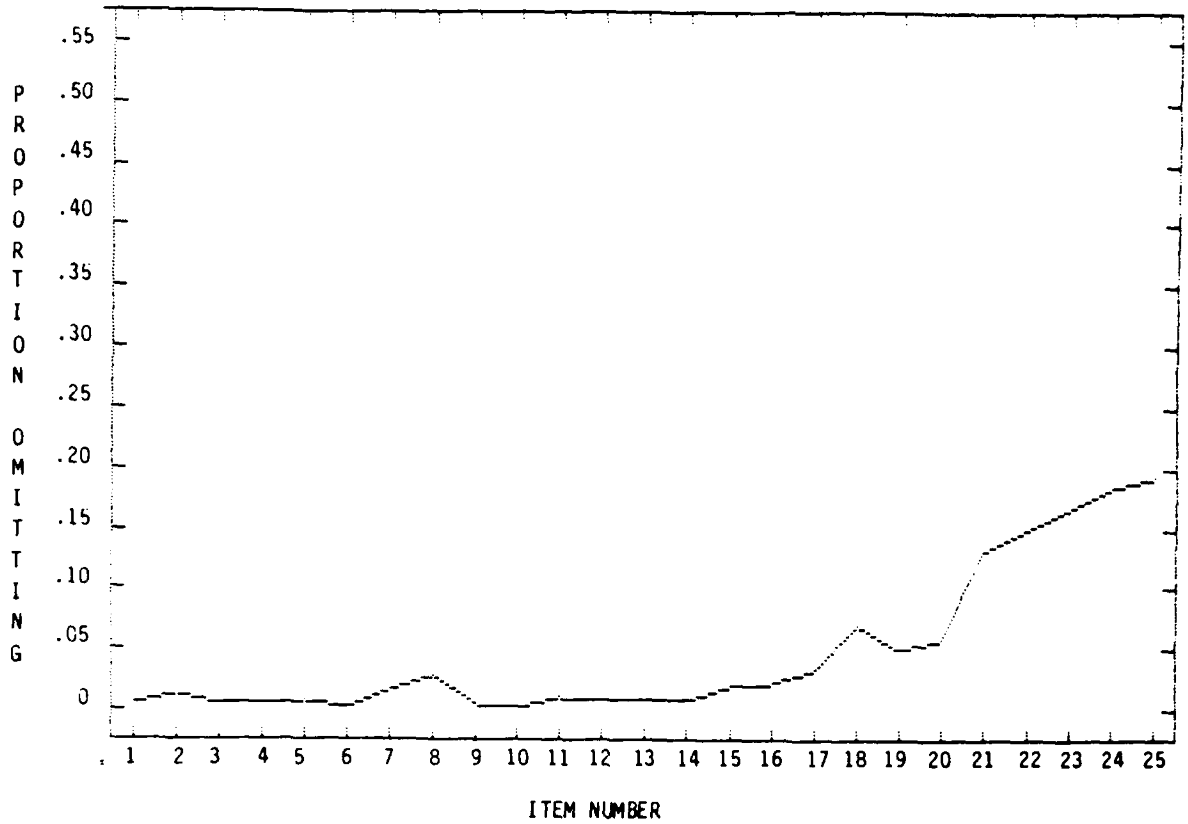


Figure B-1. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE VERBAL ANALOGIES SUBTEST

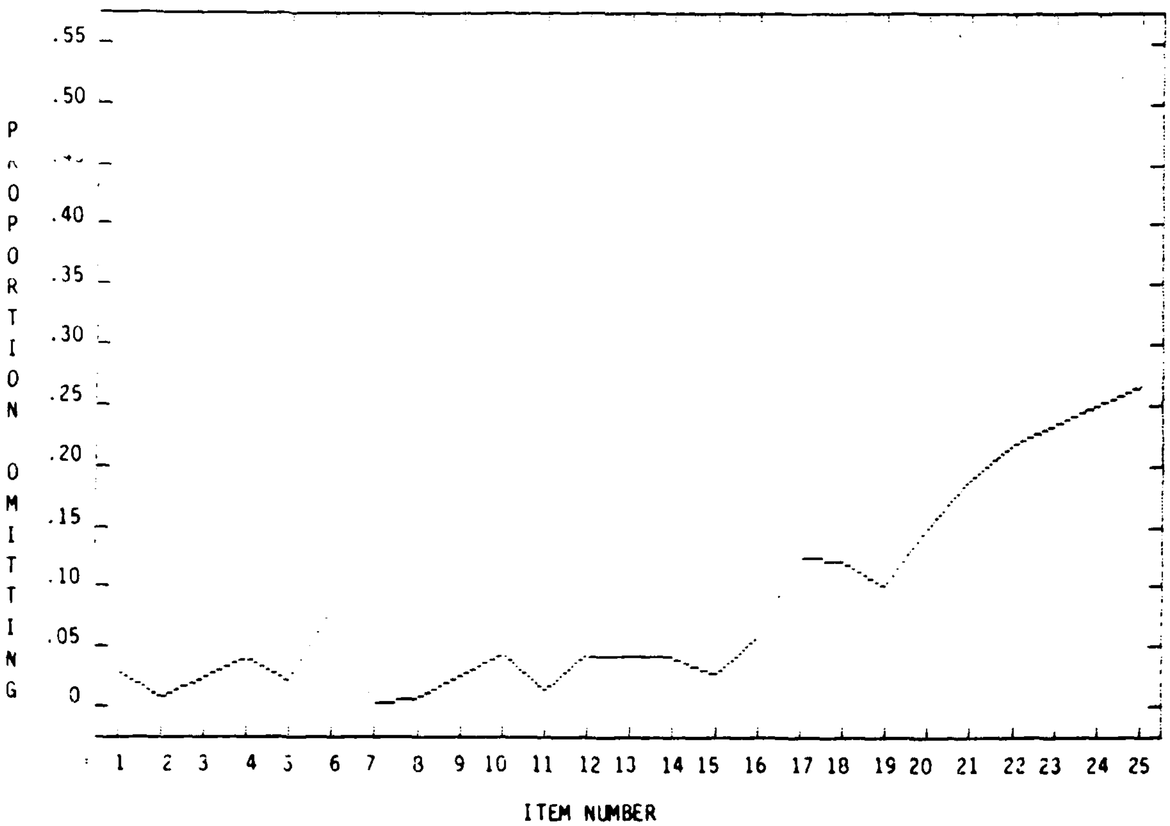


Figure B-2. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE ARITHMETIC REASONING SUBTEST

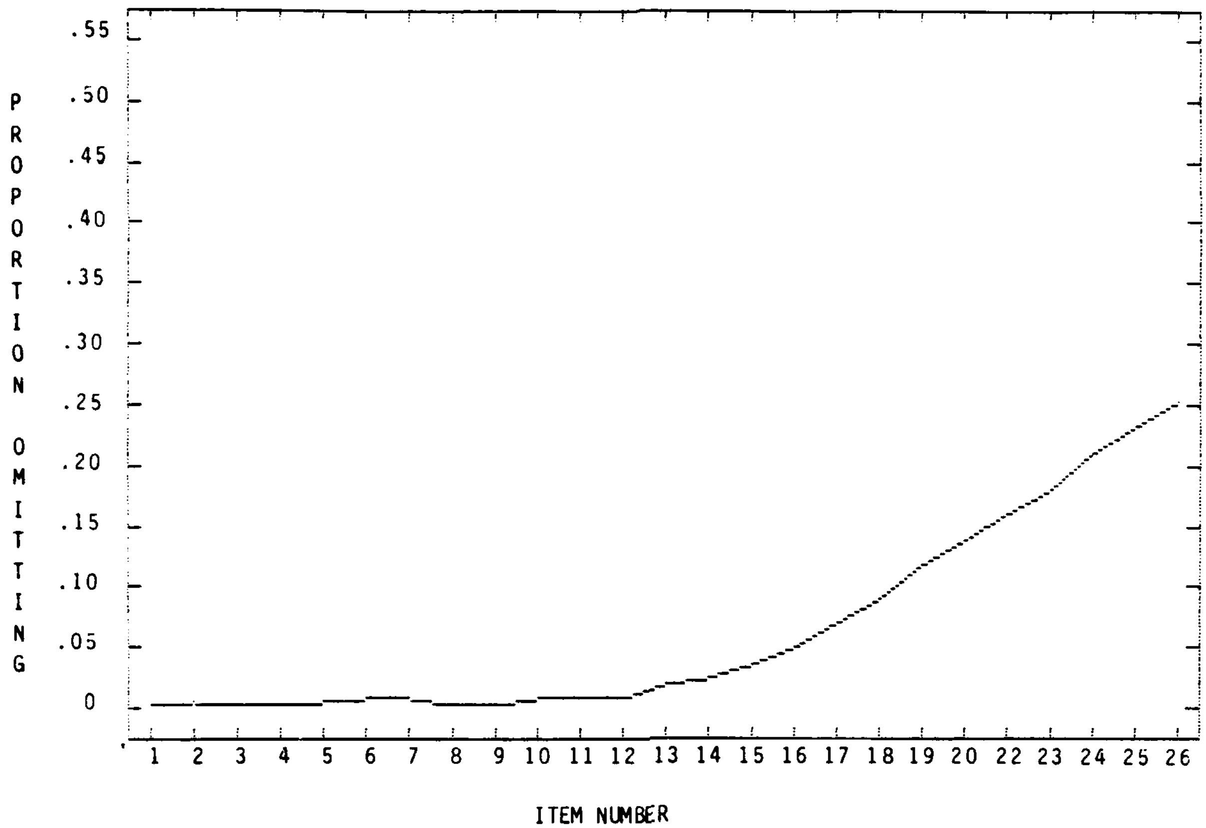


Figure B-3. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE READING COMPREHENSION SUBTEST

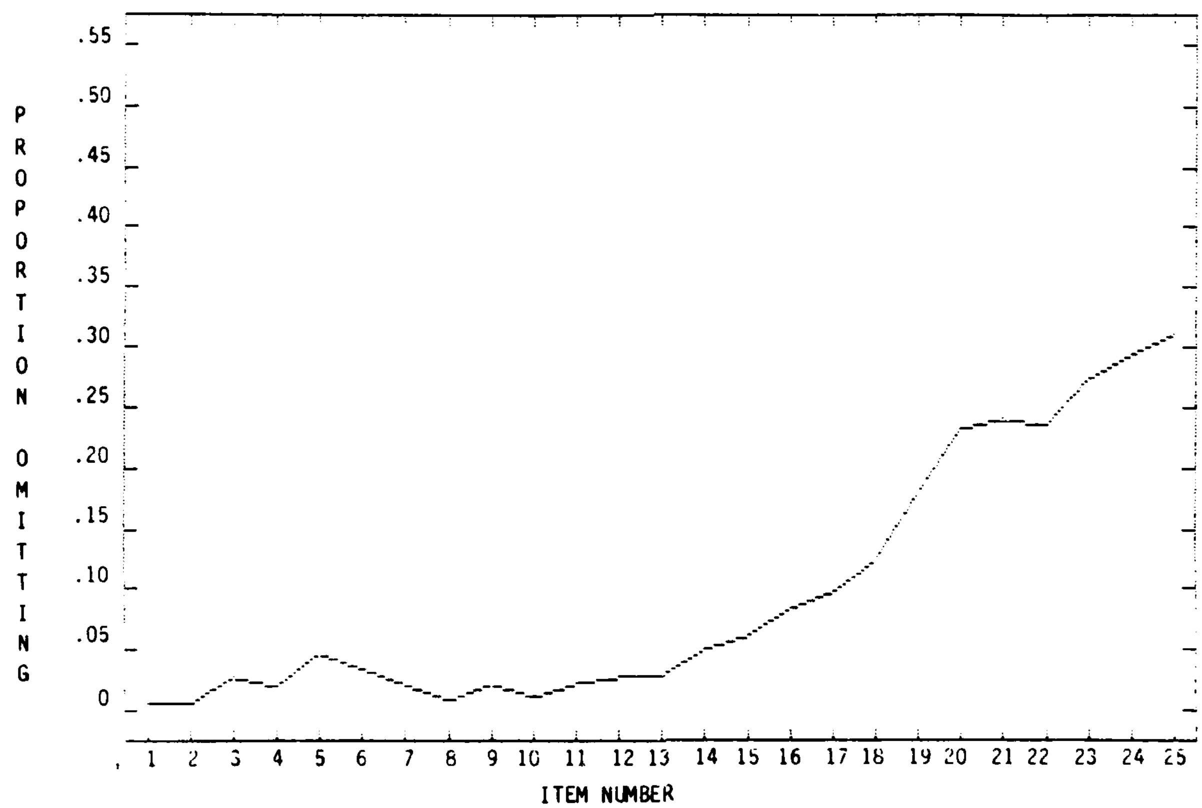


Figure B-4. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE DATA INTERPRETATION SUBTEST

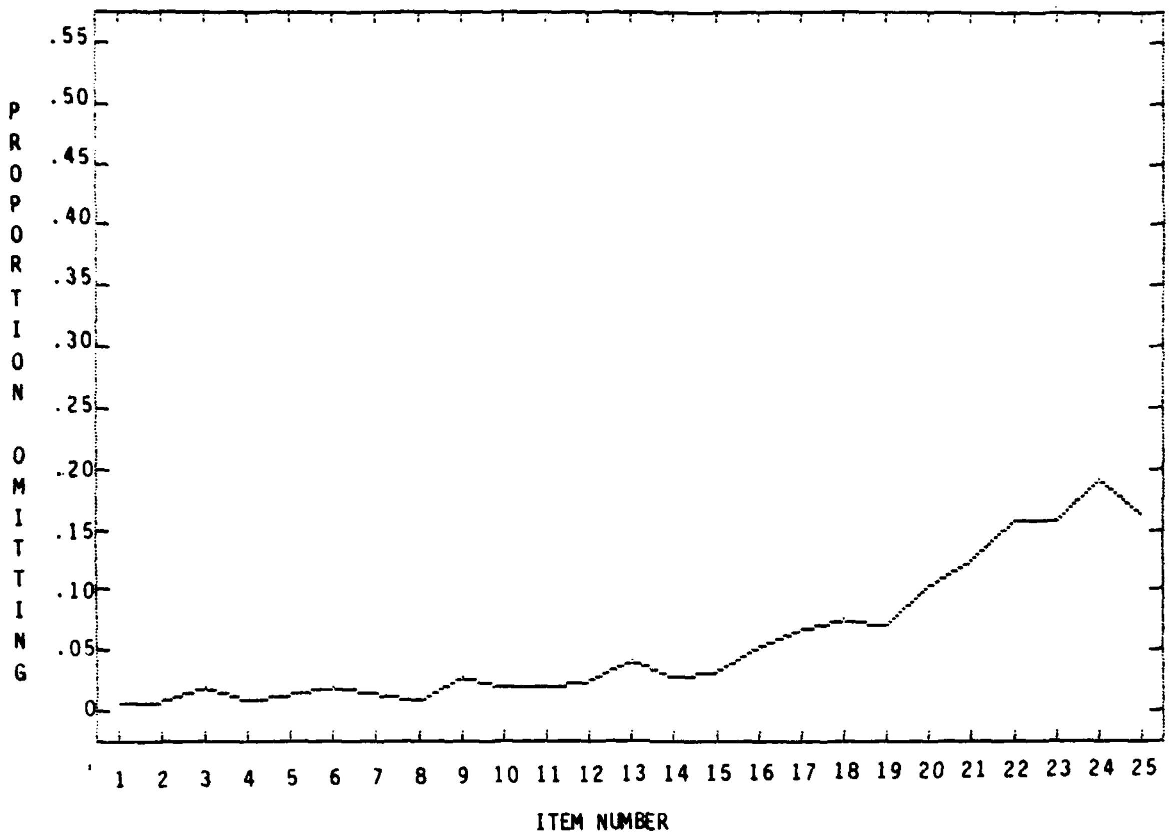


Figure B-5. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE WORD KNOWLEDGE SUBTEST

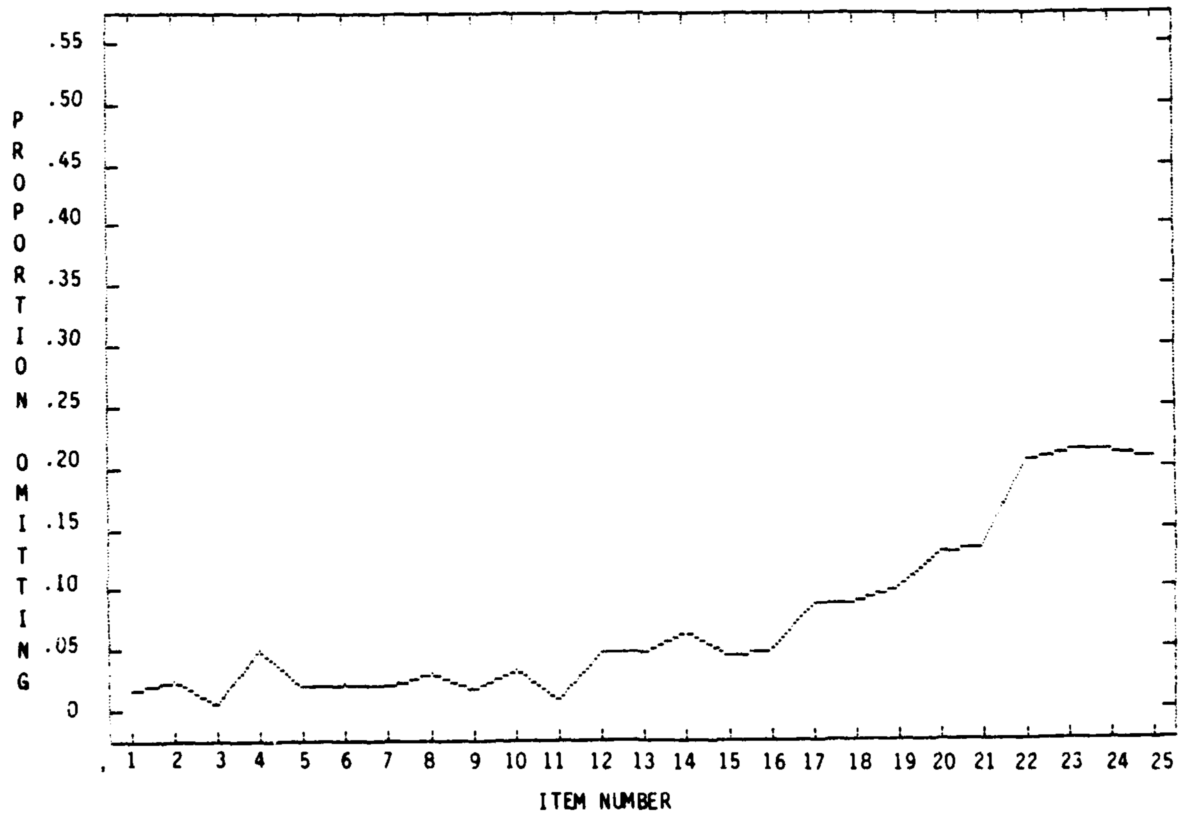


Figure B-6. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE MATH KNOWLEDGE SUBTEST

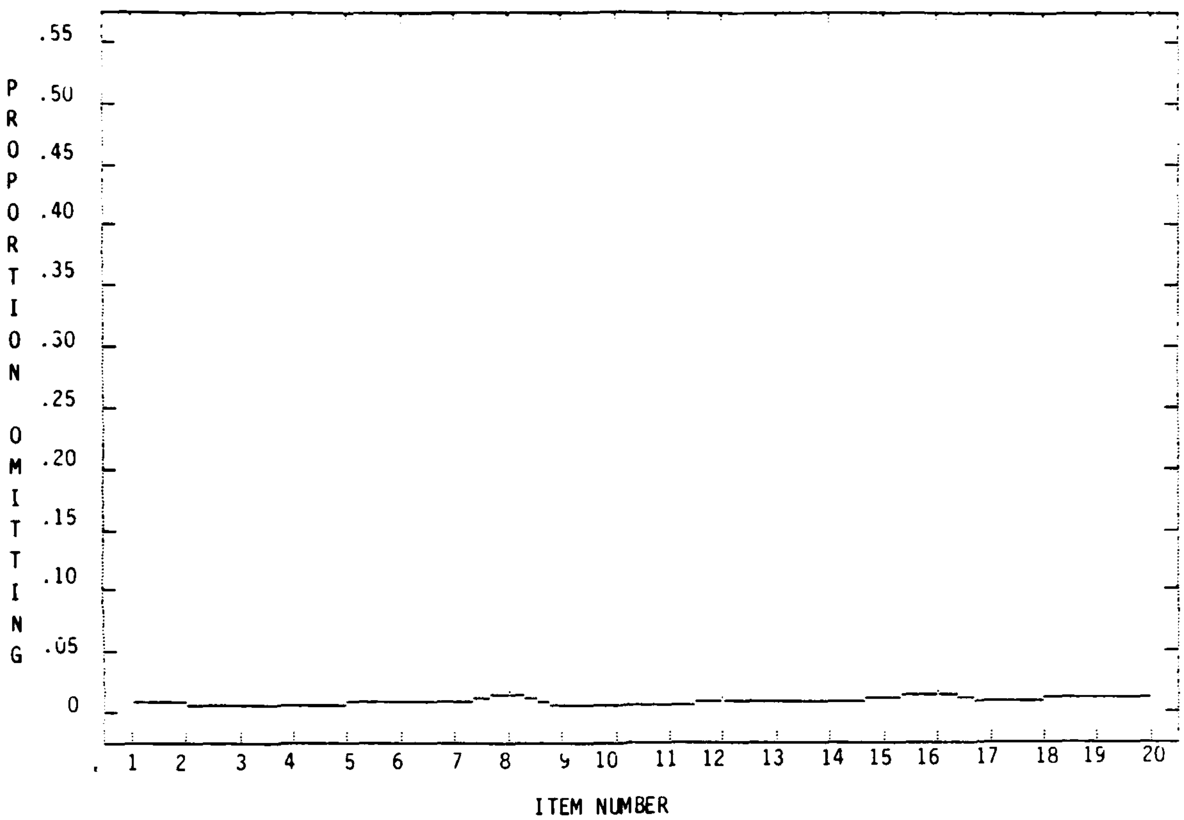


FIGURE B-7. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE MECHANICAL COMPREHENSION SUBTEST

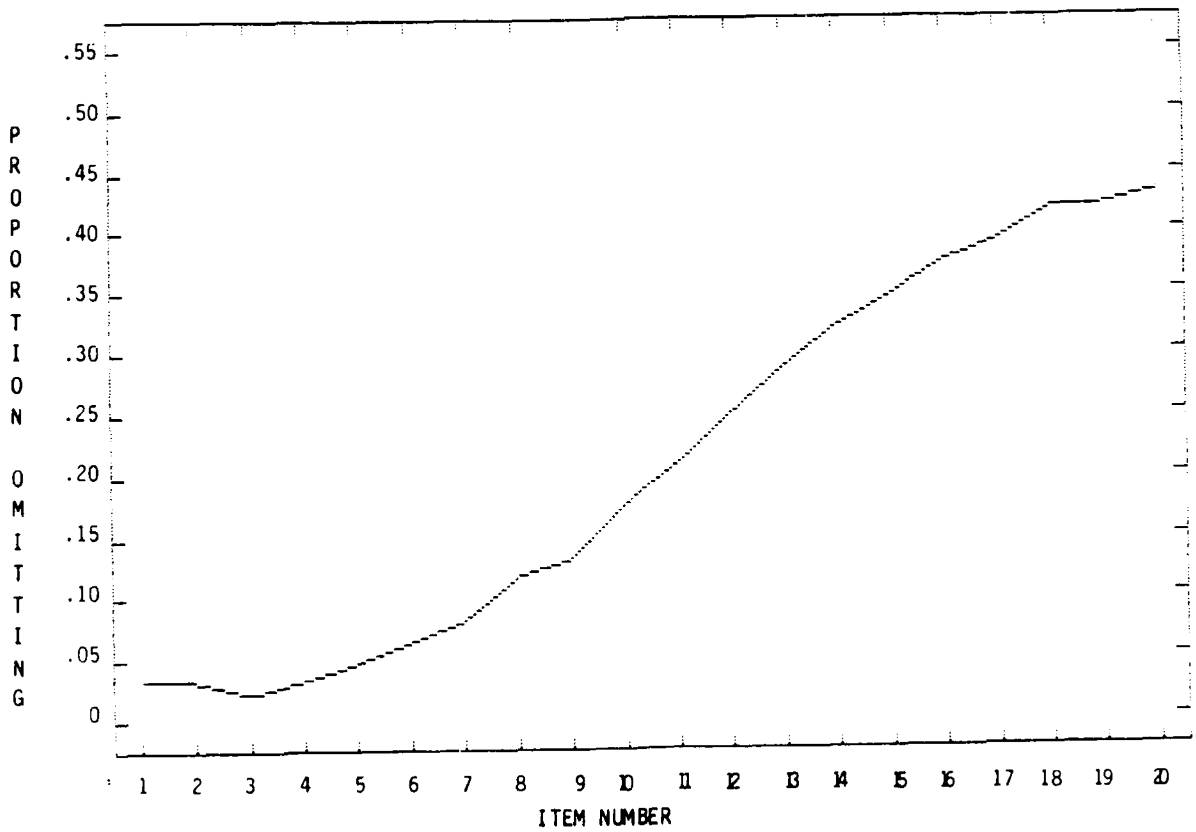


FIGURE B-8. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE ELECTRICAL MAZE SUBTEST

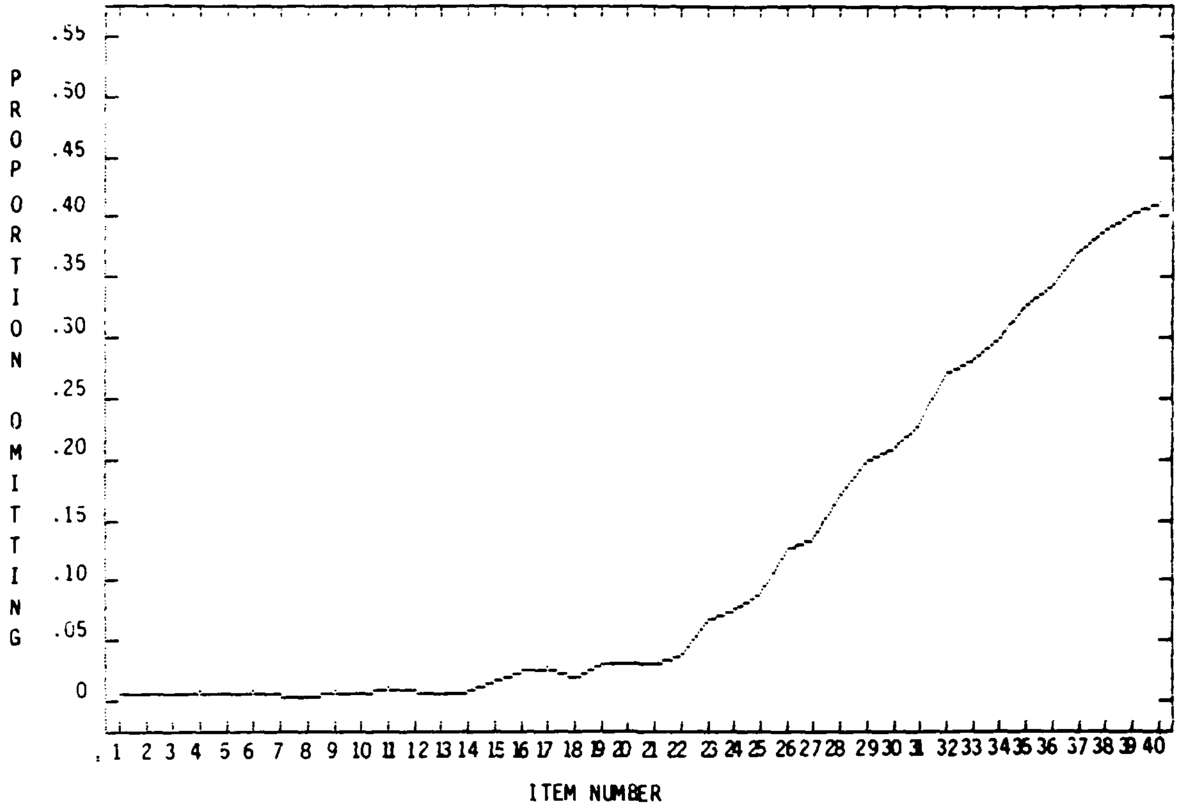


FIGURE B-9. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE SCALE READING SUBTEST

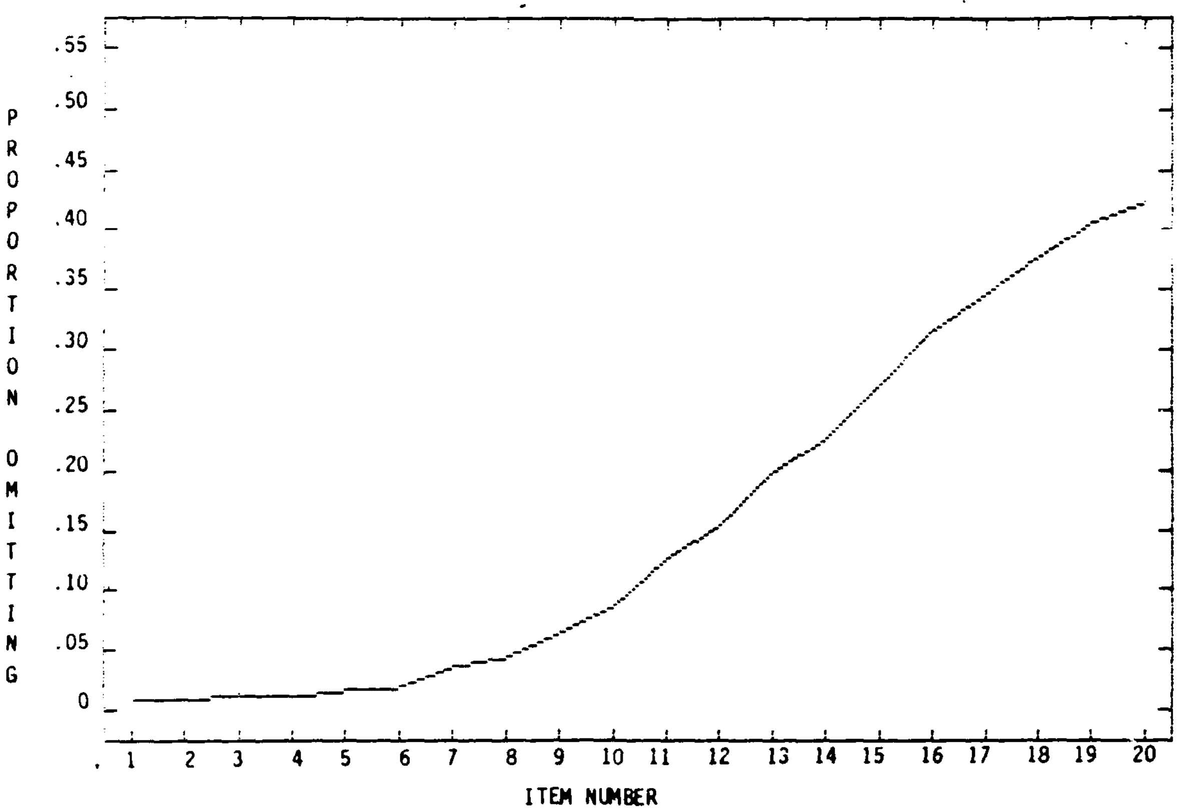


FIGURE B-10. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE INSTRUMENT COMPREHENSION SUBTEST

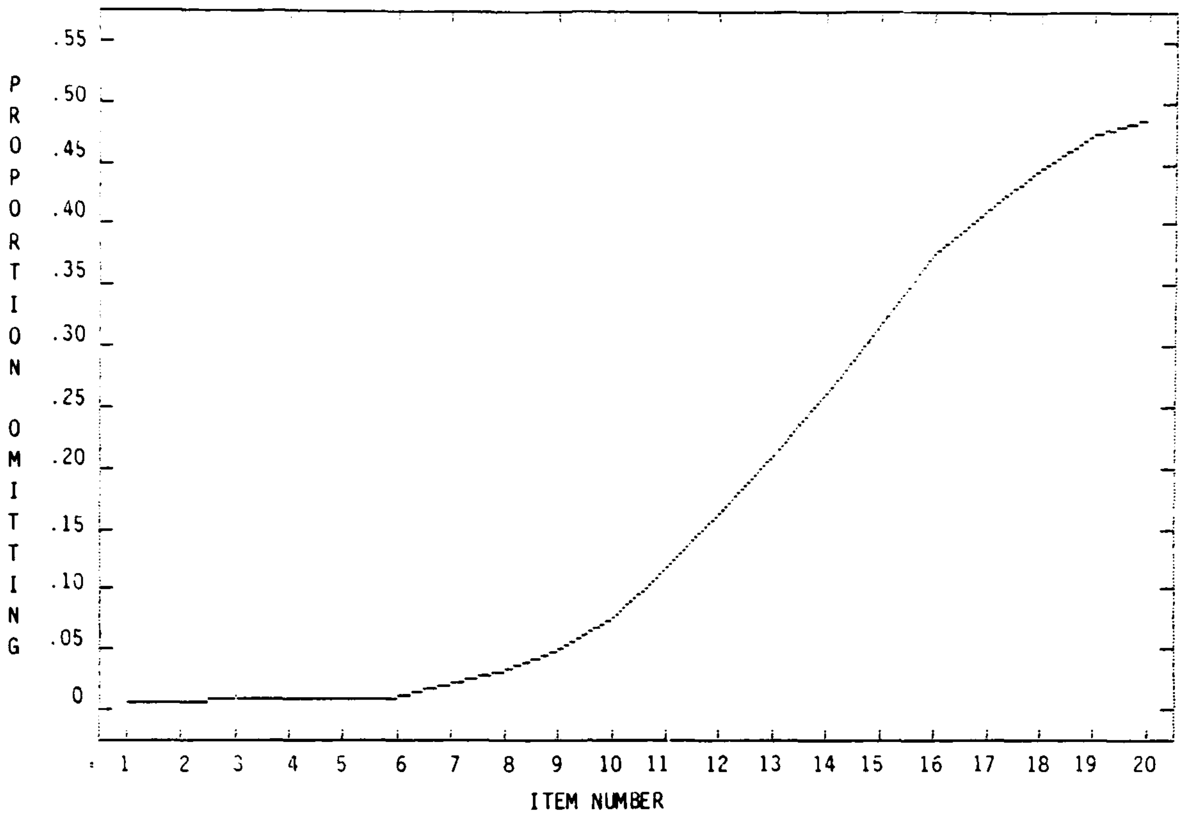


FIGURE B-11. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE BLOCK COUNTING SUBTEST

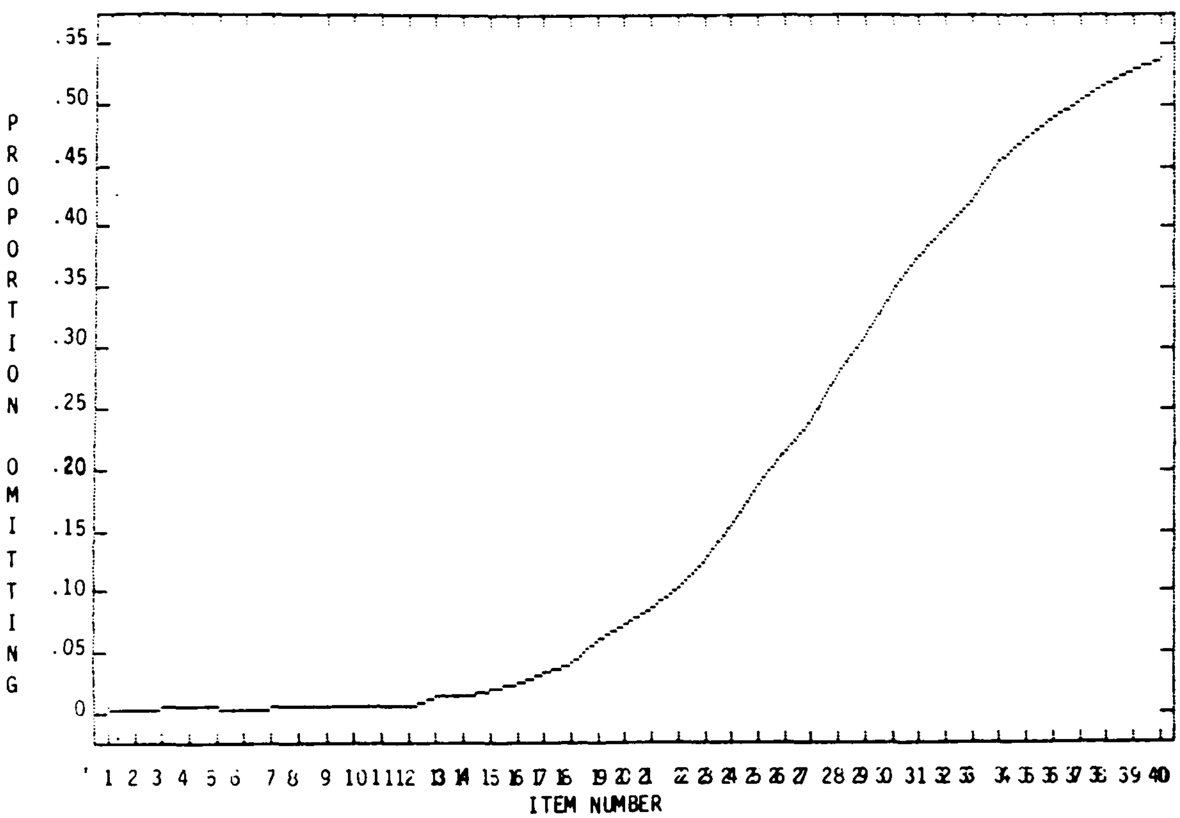


FIGURE B-12. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE TABLE READING SUBTEST

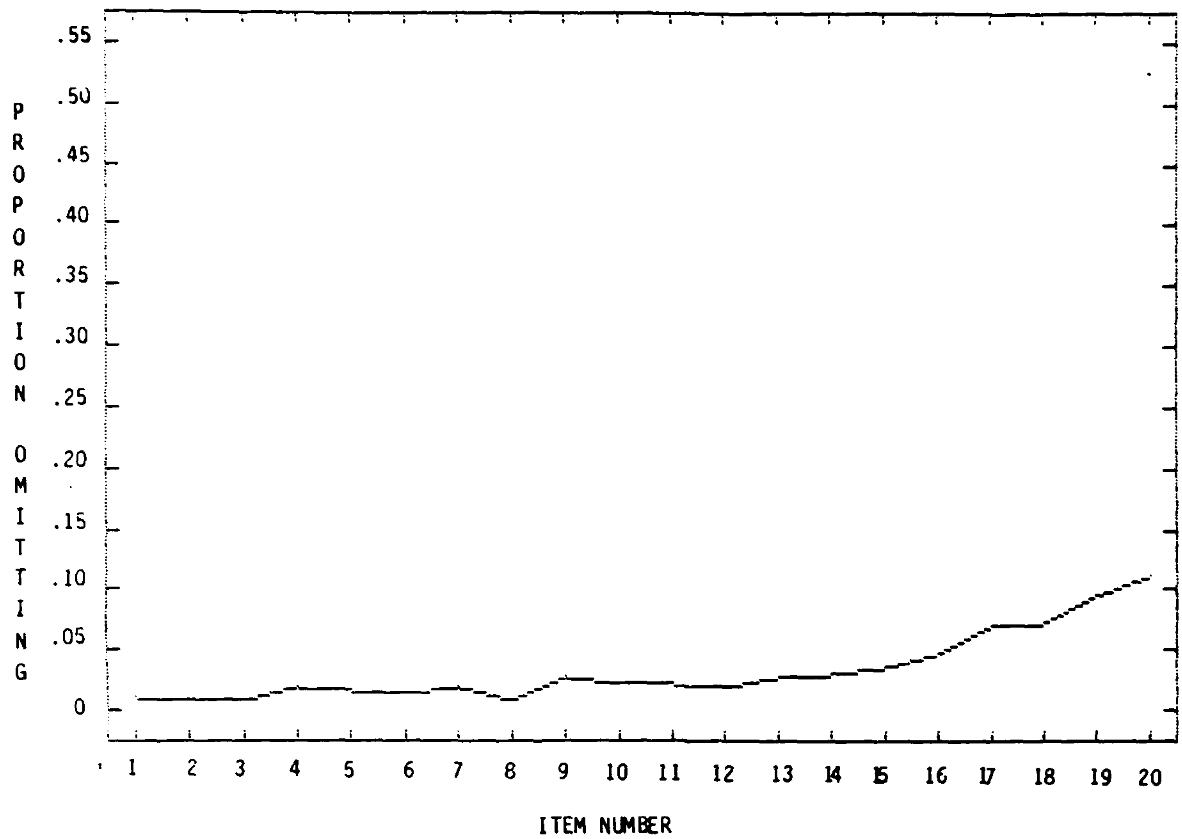


FIGURE B-13. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE AVIATION INFORMATION SUBTEST

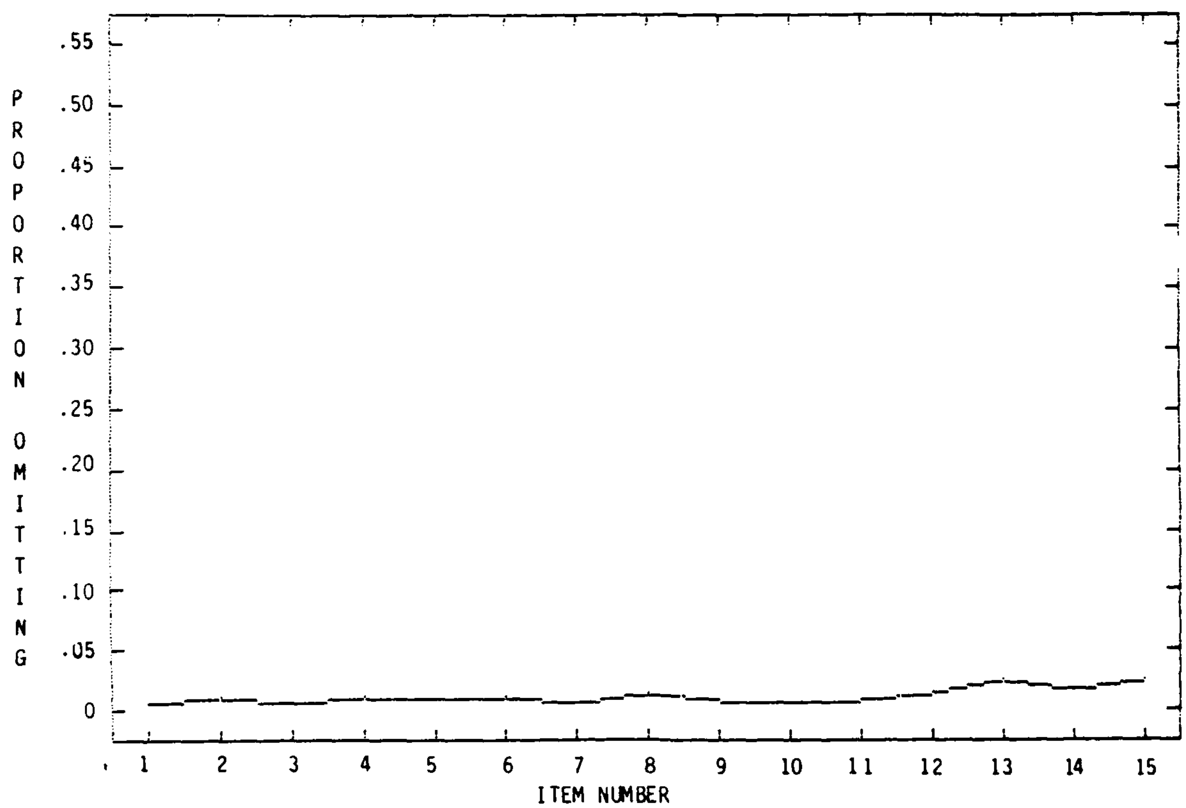


FIGURE B-14. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE ROTATED BLOCKS SUBTEST

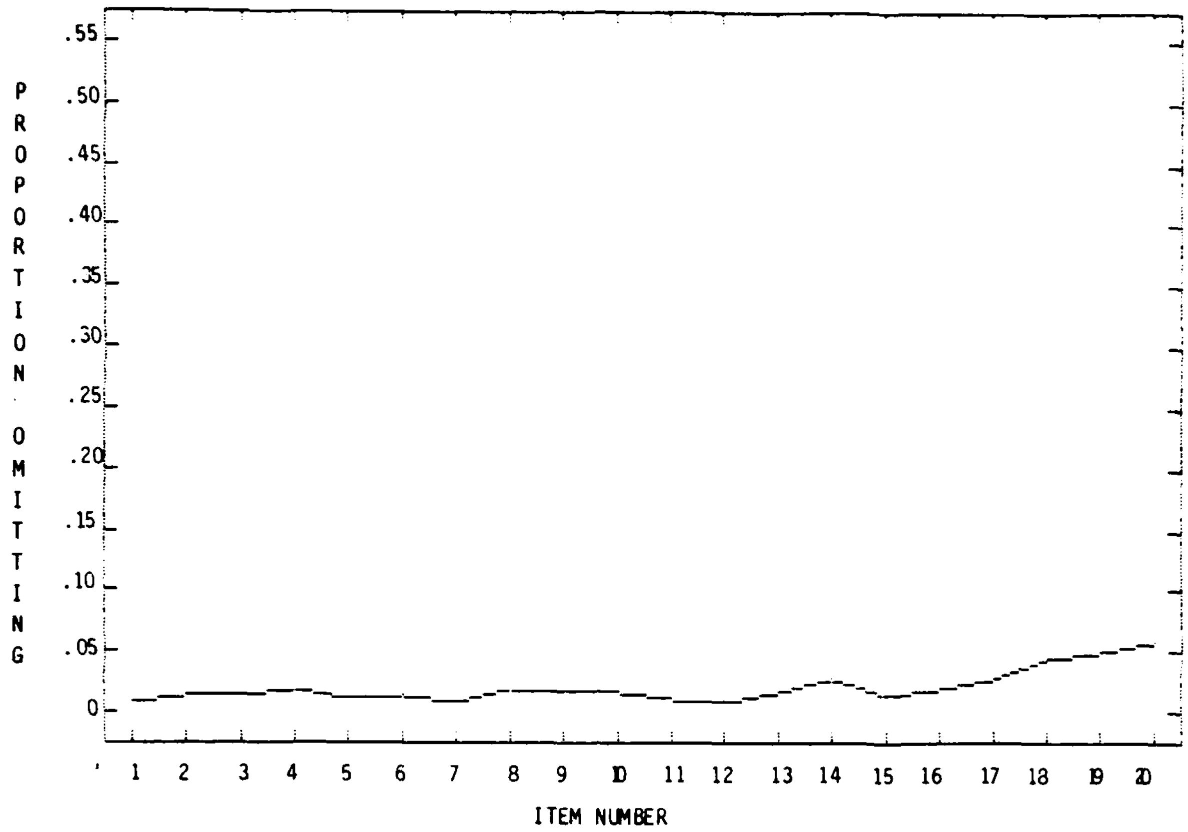


FIGURE B-15. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE GENERAL SCIENCE SUBTEST

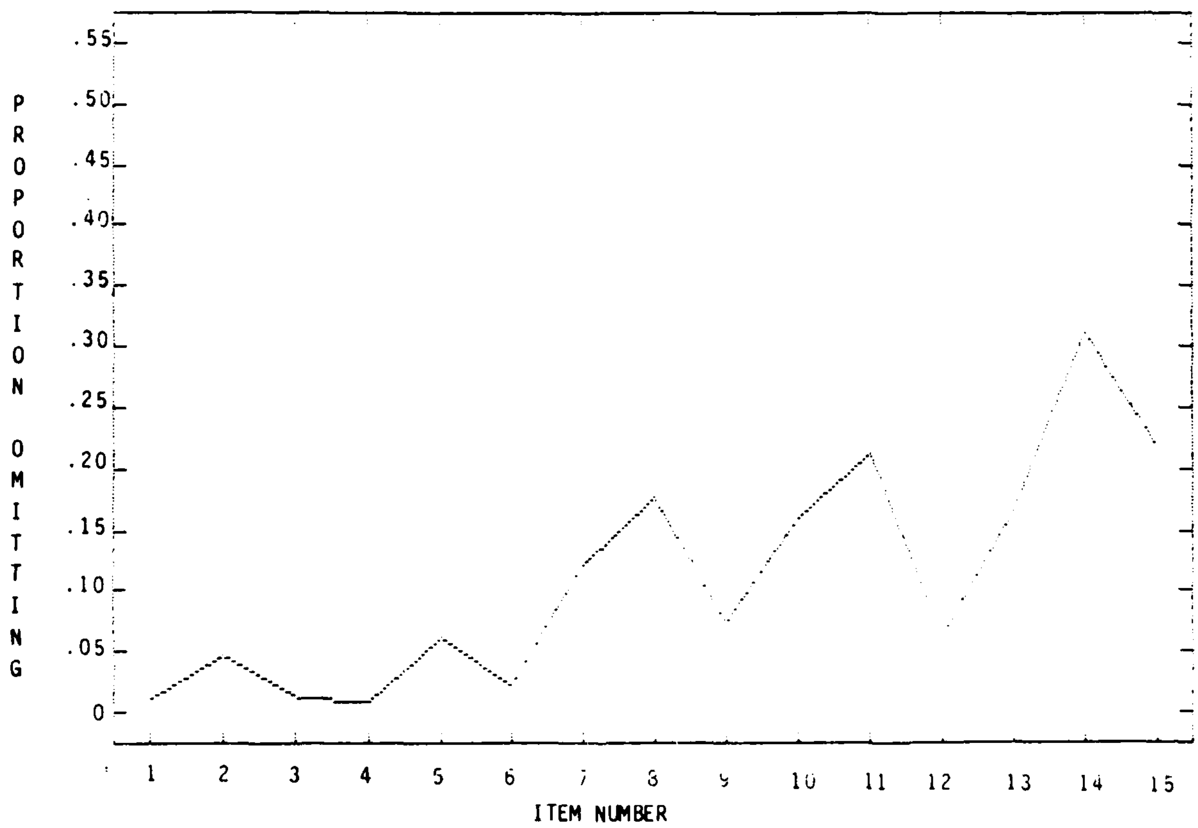


FIGURE B-16. PROPORTION OF EXAMINEES OMITTING EACH ITEM IN THE HIDDEN FIGURES SUBTEST

APPENDIX C. TEST INFORMATION CURVES FOR THE SUBTESTS

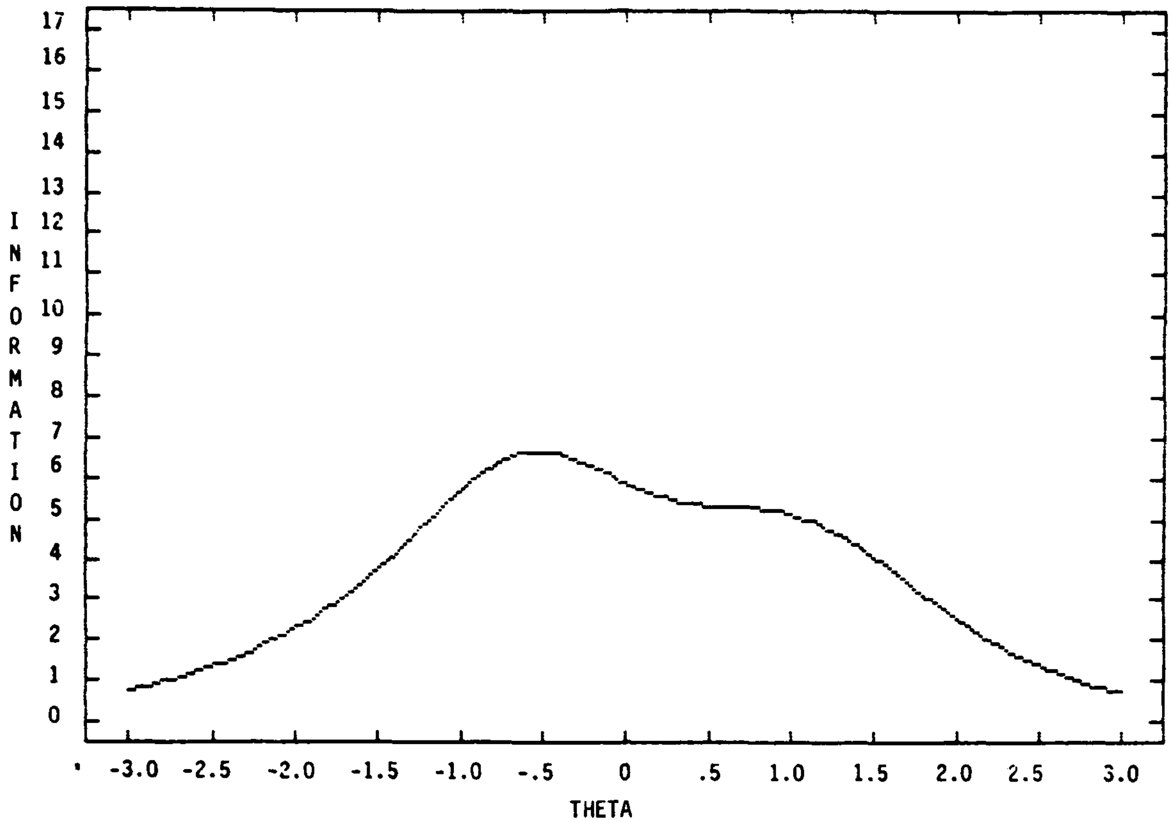


Figure C-1. TEST INFORMATION CURVE FOR VERBAL ANALOGIES SUBTEST

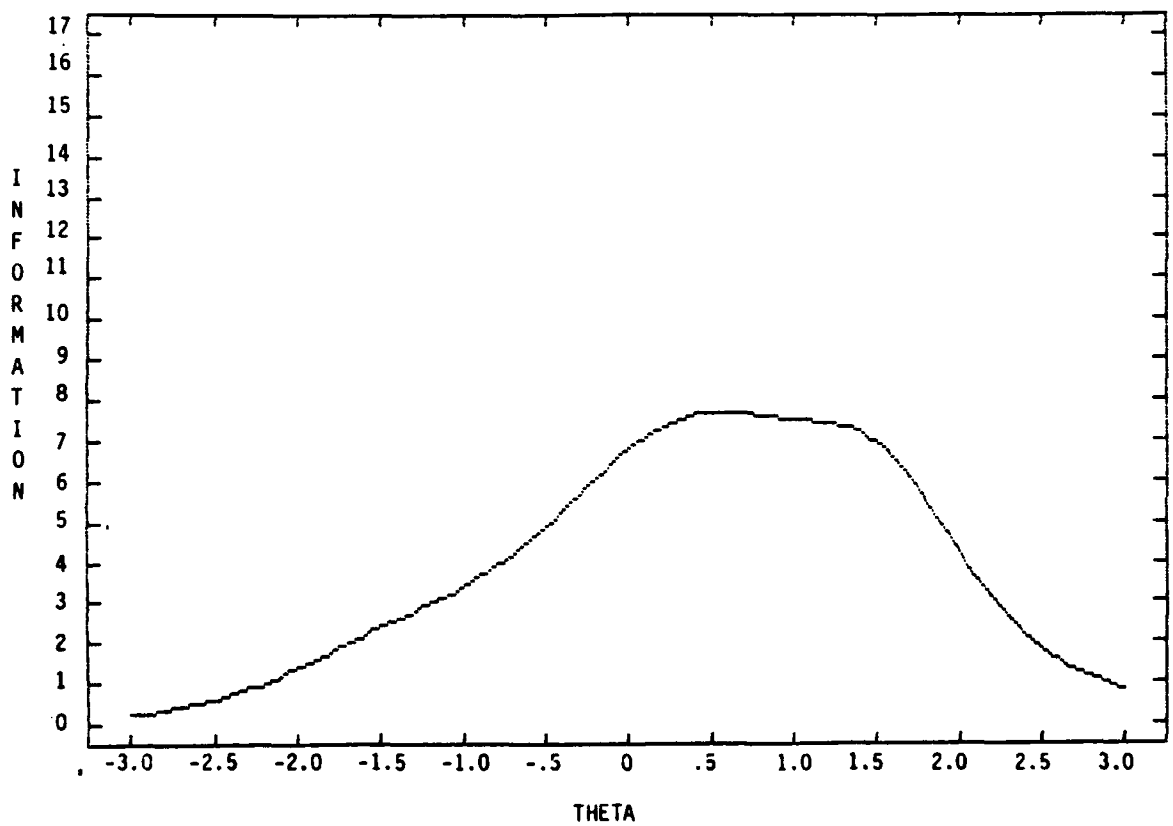


Figure C-2. TEST INFORMATION CURVE FOR ARITHMETIC REASONING SUBTEST

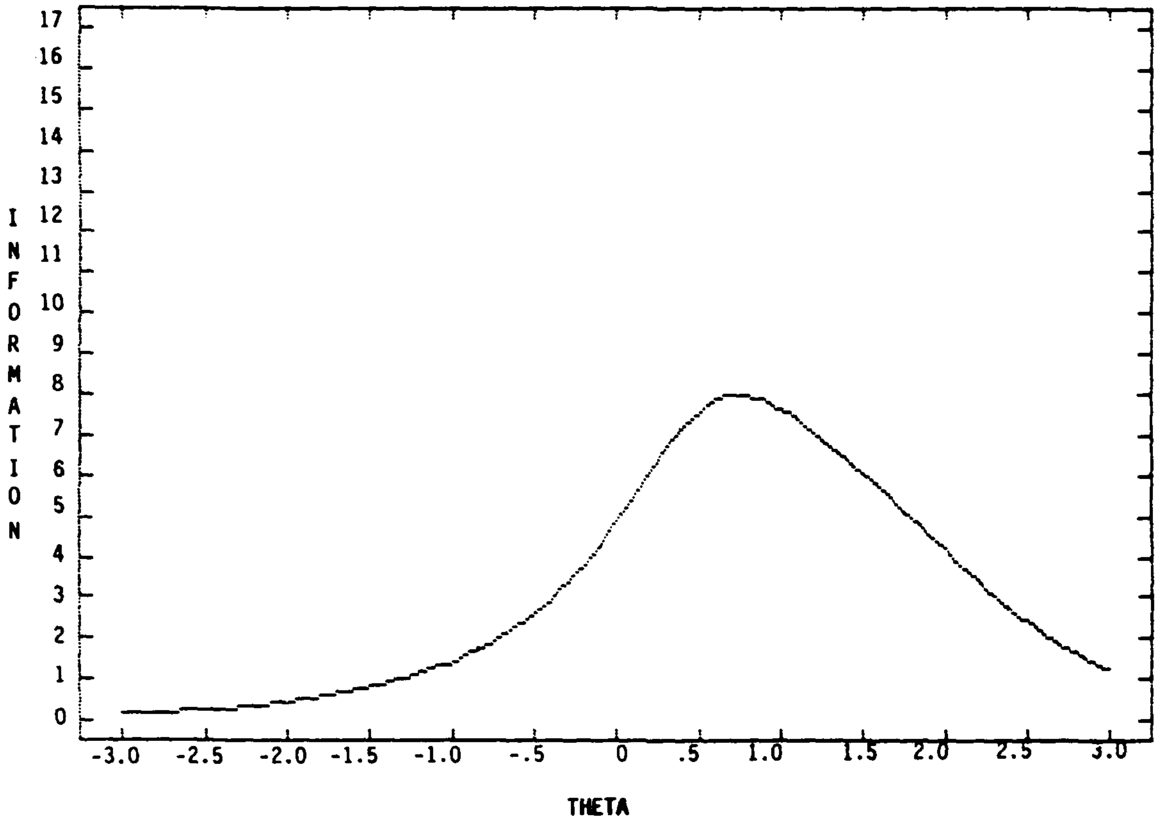


Figure C-3. TEST INFORMATION CURVE FOR AVIATION INFORMATION SUBTEST

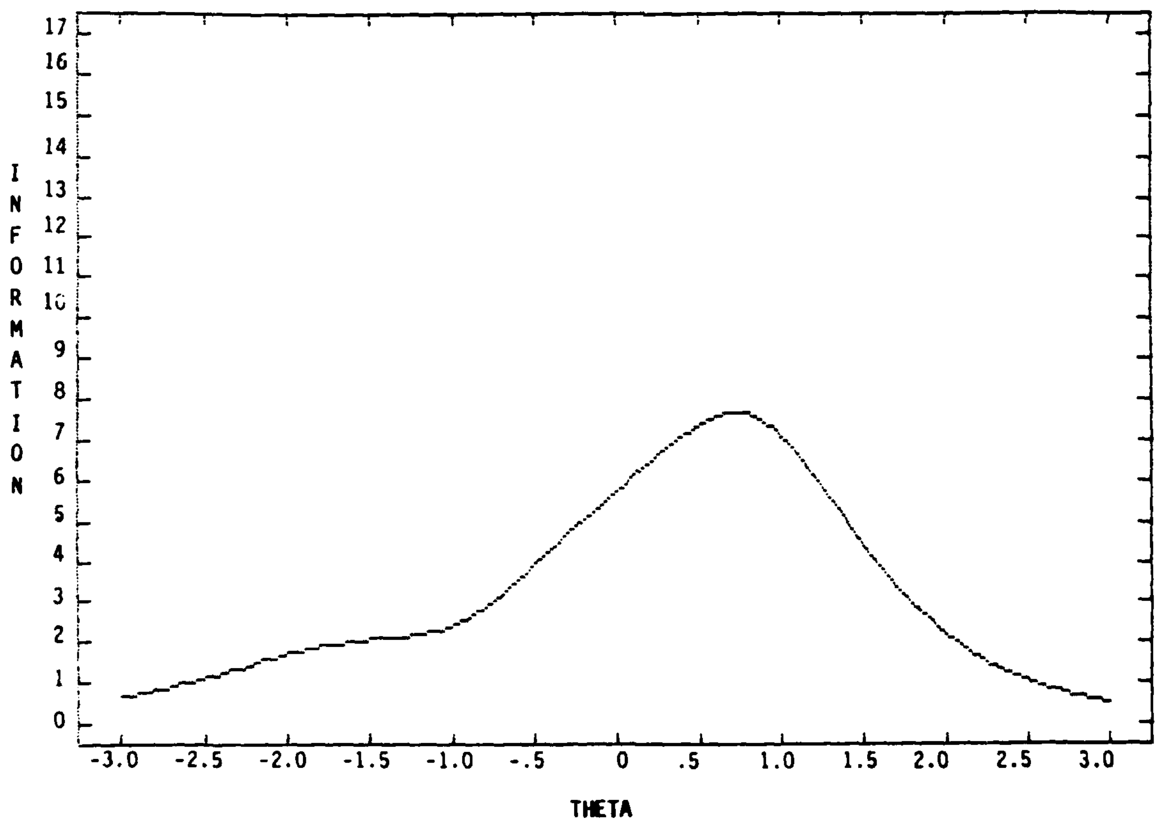


Figure C-4. TEST INFORMATION CURVE FOR ROTATED BLOCKS SUBTEST

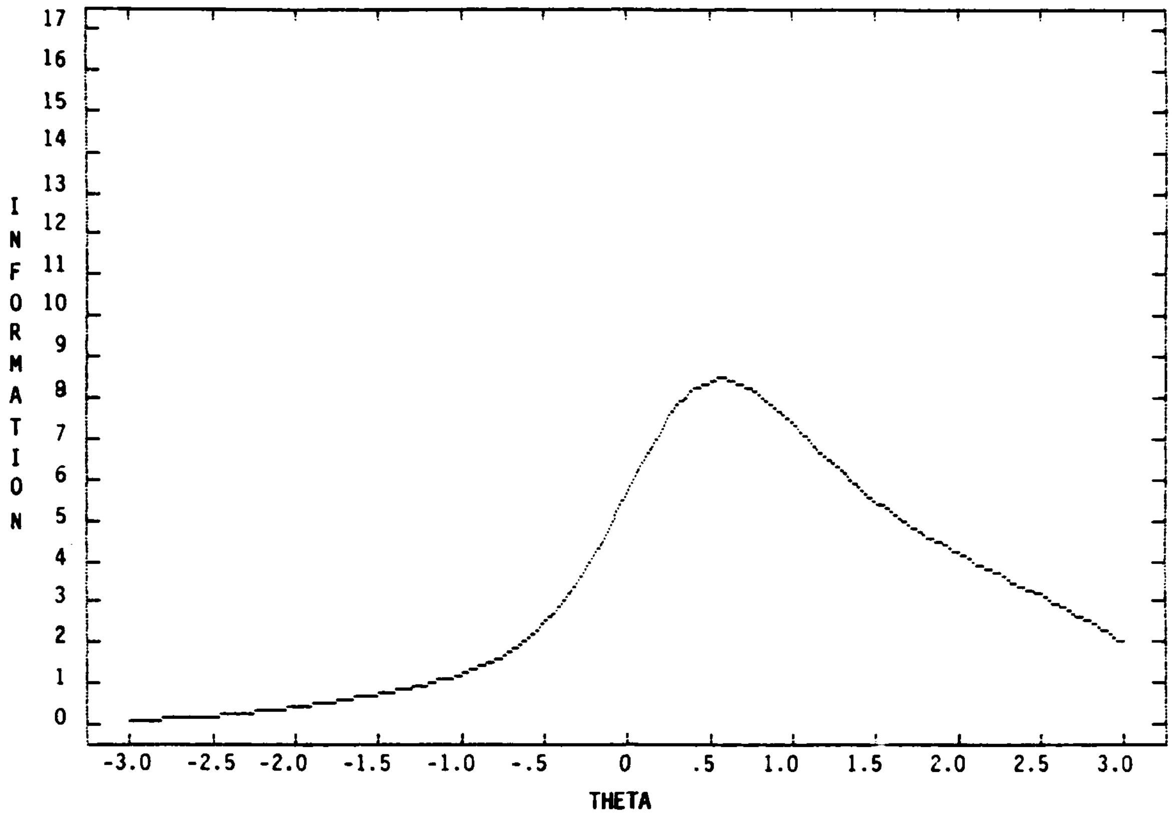


Figure C-5. TEST INFORMATION CURVE FOR GENERAL SCIENCE SUBTEST

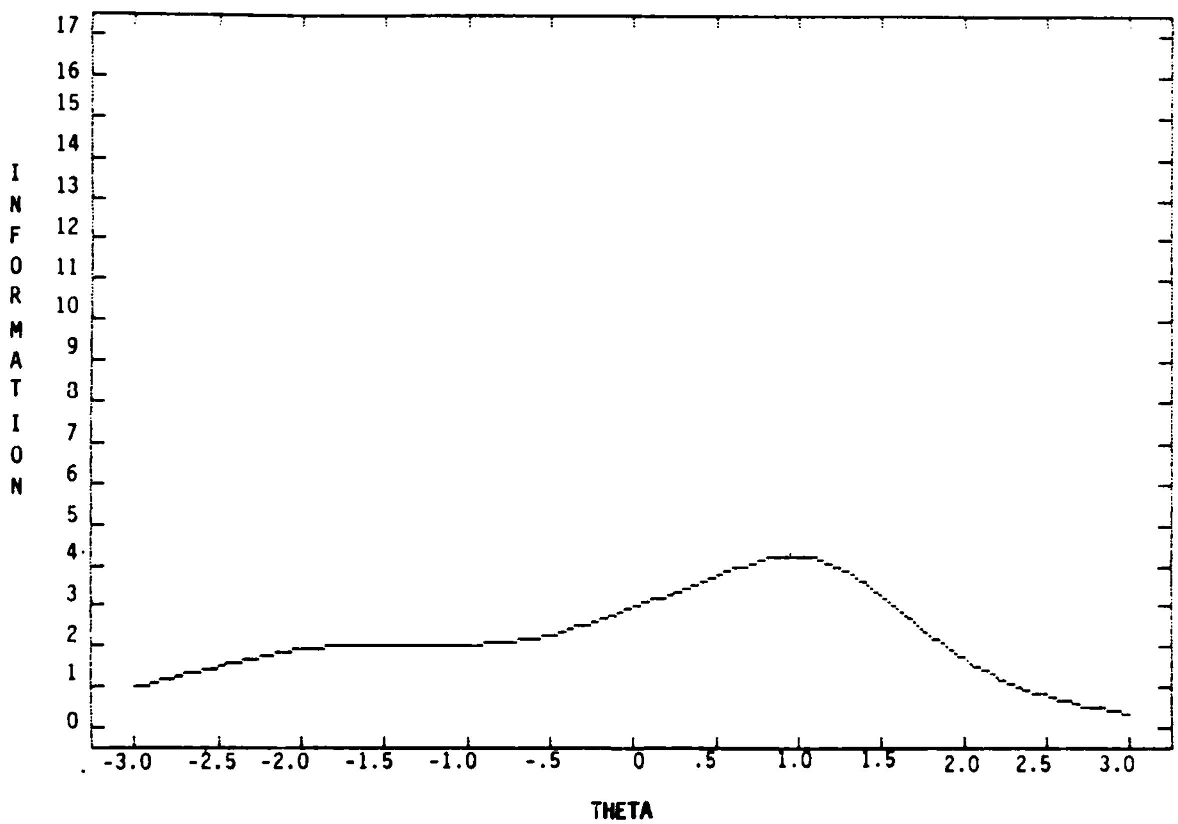


Figure C-6. TEST INFORMATION CURVE FOR HIDDEN FIGURES SUBTEST

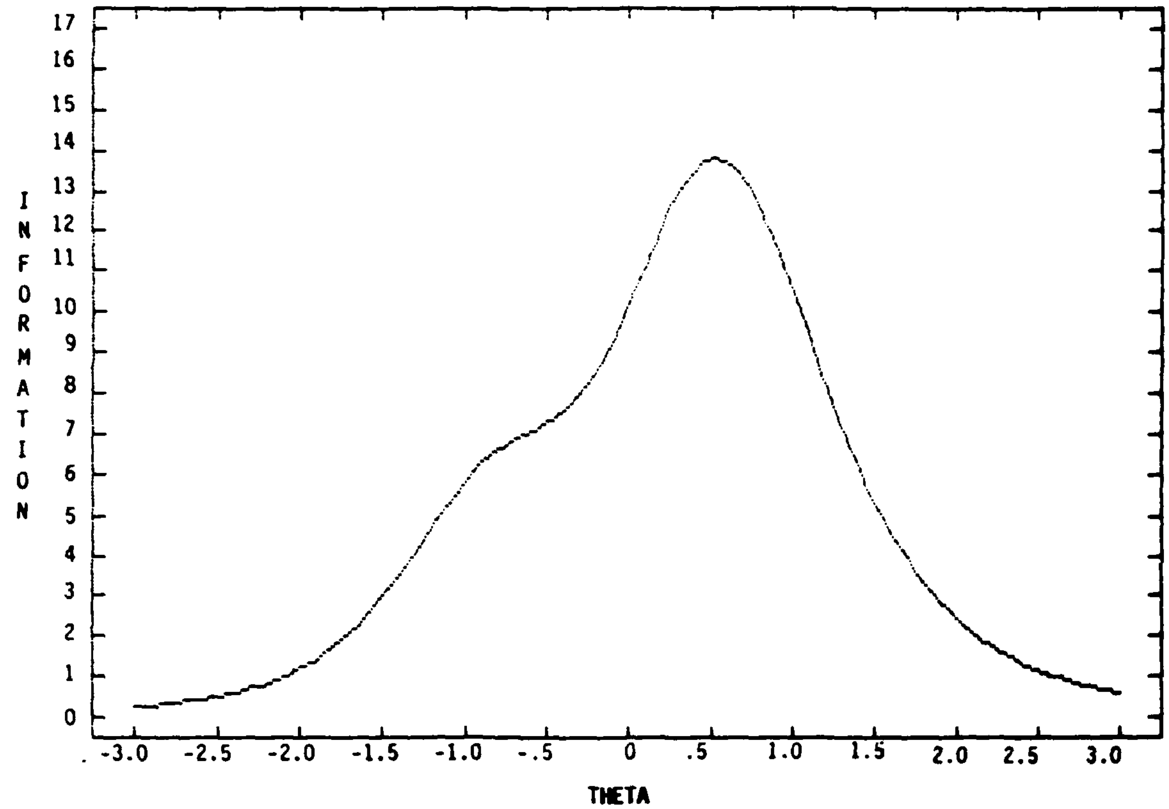


Figure C-7. TEST INFORMATION CURVE FOR WORD KNOWLEDGE SUBTEST

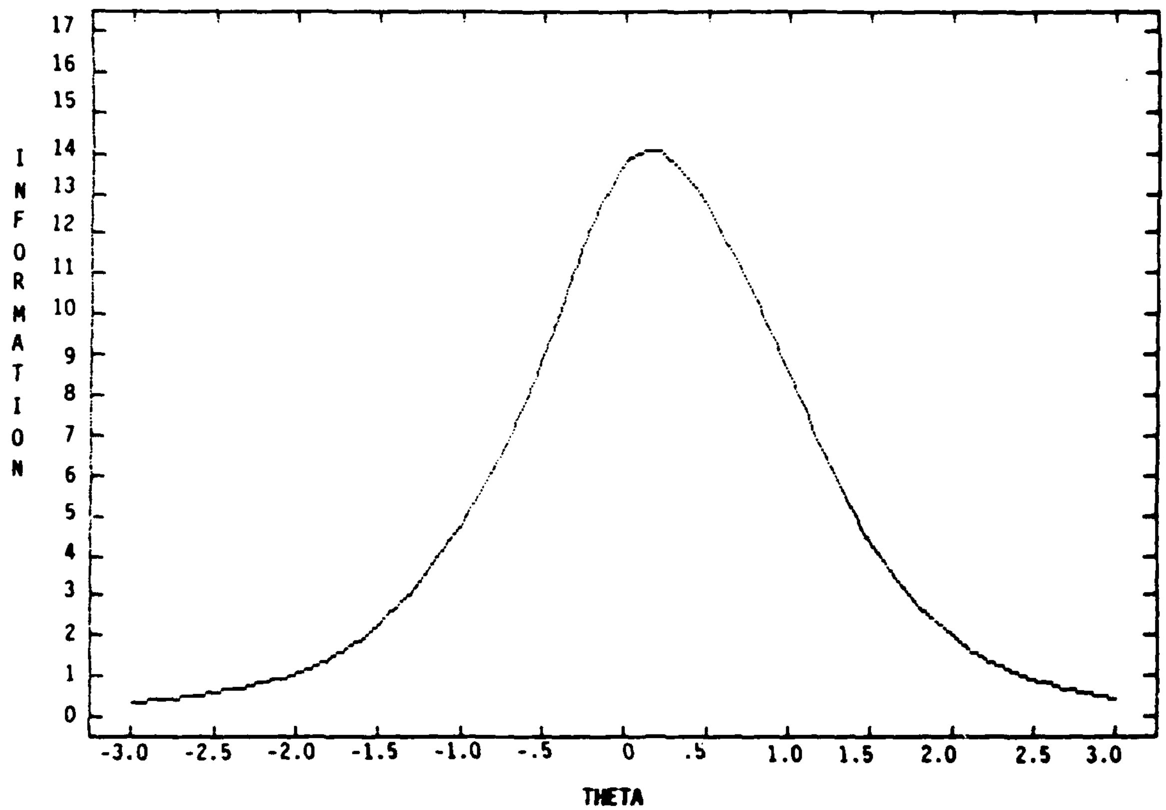


Figure C-8. TEST INFORMATION CURVE FOR MATH KNOWLEDGE SUBTEST

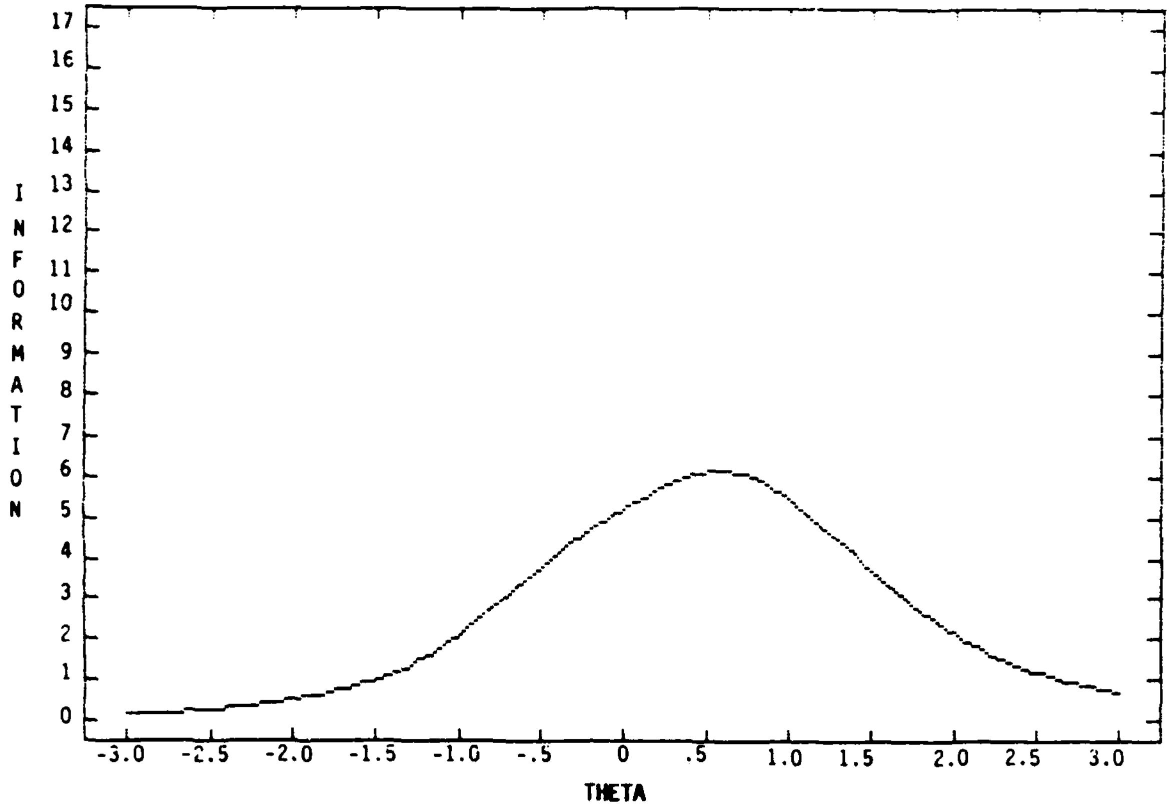


Figure C-9. TEST INFORMATION CURVE FOR MECHANICAL COMPREHENSION SUBTEST

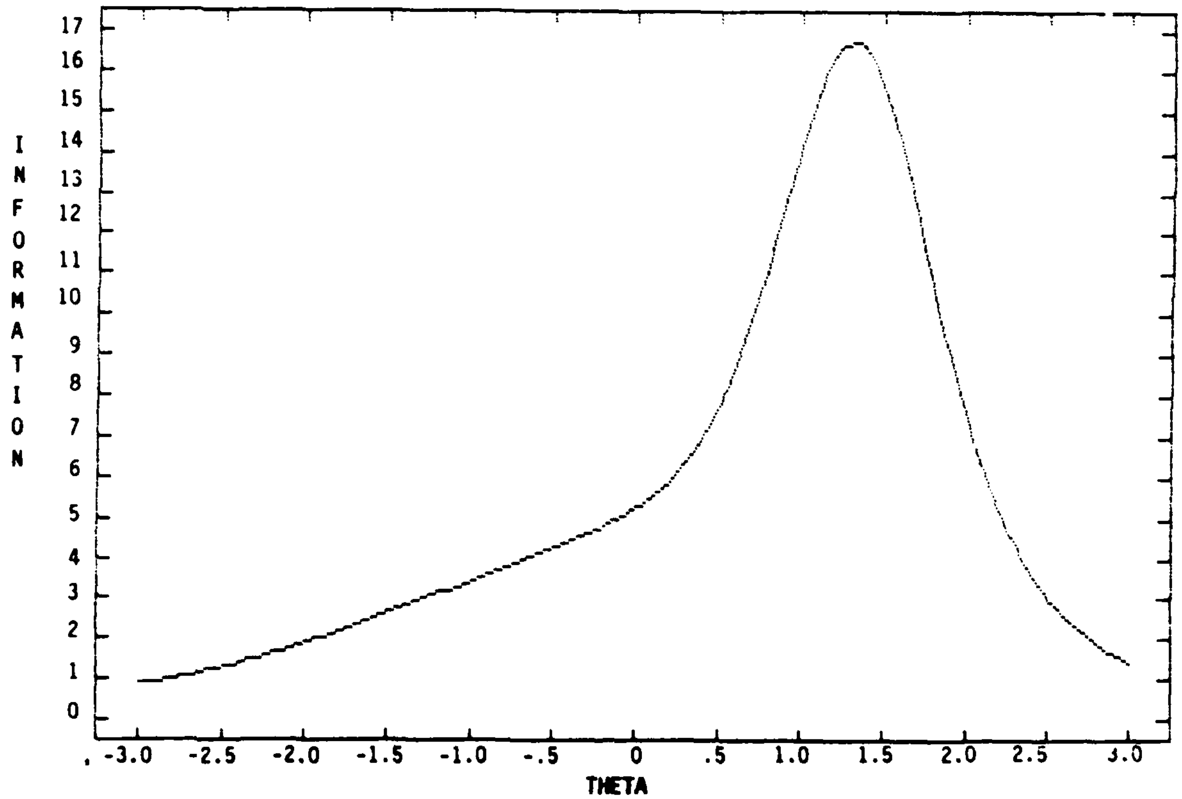


Figure C-10. TEST INFORMATION CURVE FOR SCALE READING SUBTEST

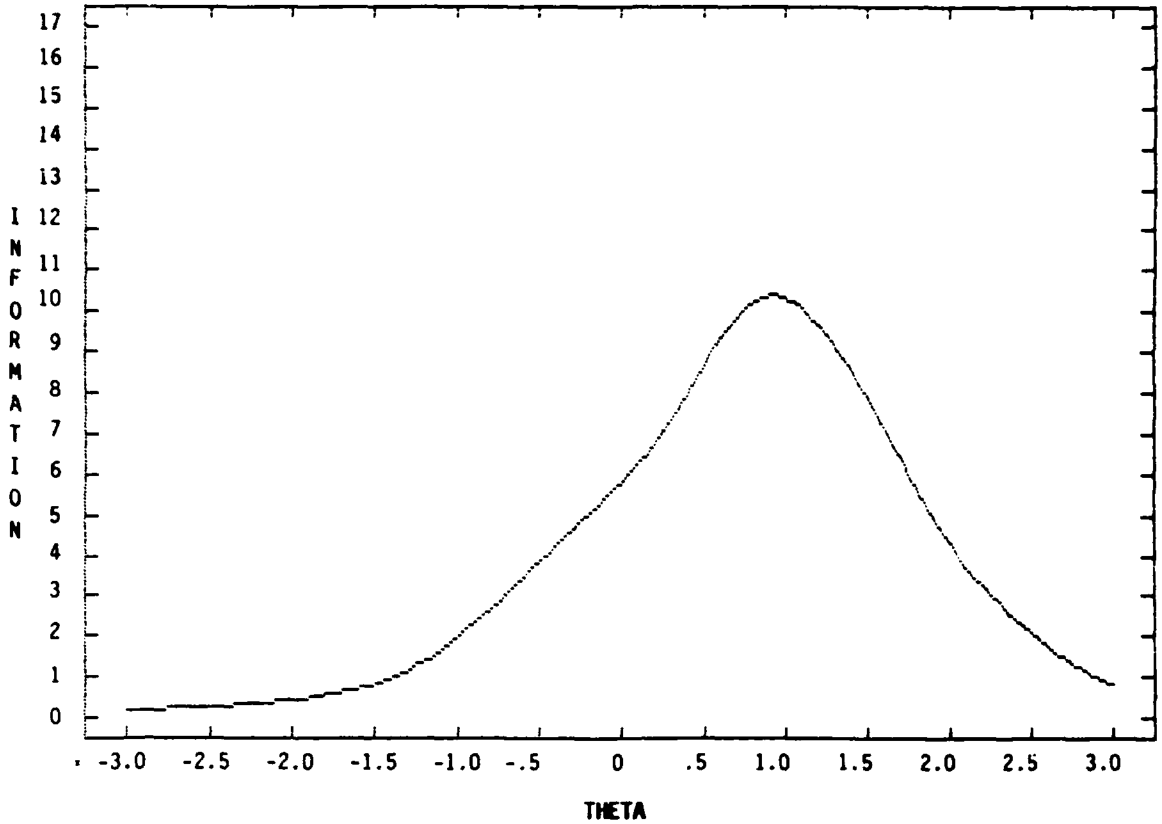


Figure C-11. TEST INFORMATION CURVE FOR INSTRUMENT COMPREHENSION SUBTEST

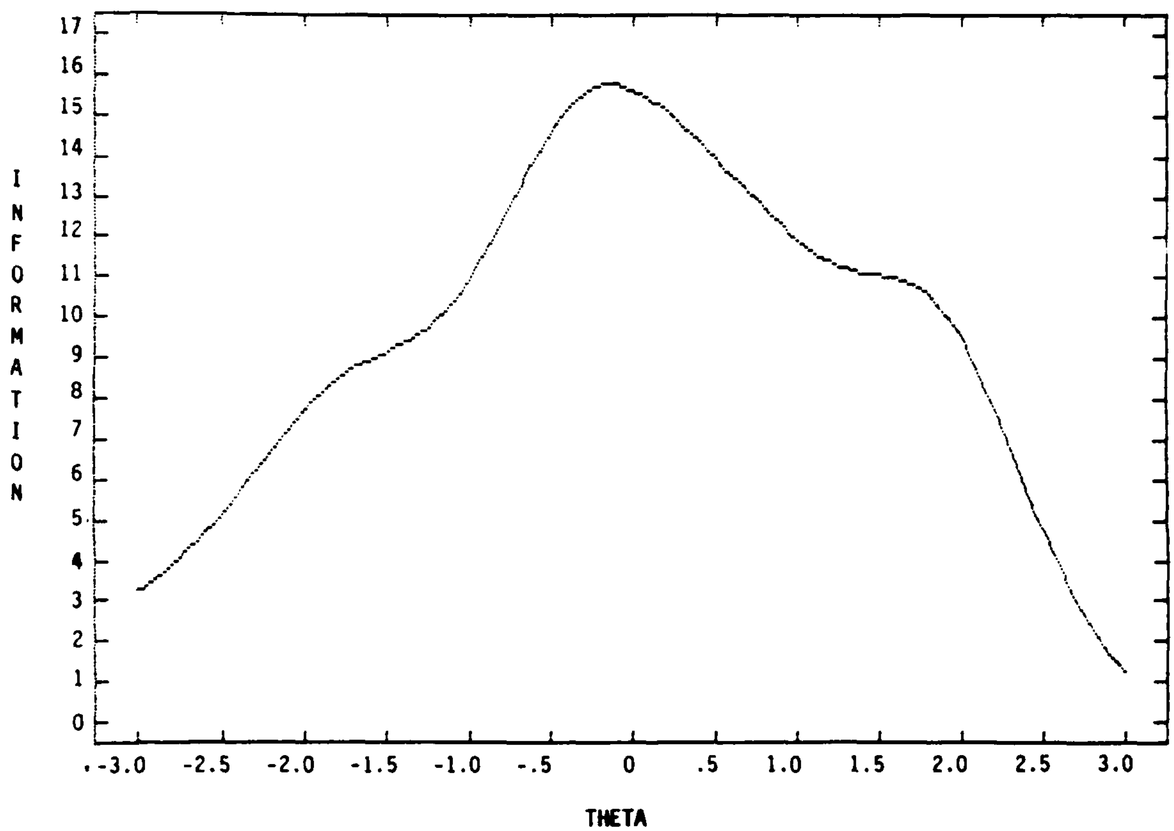


Figure C-12. TEST INFORMATION CURVE FOR TABLE READING SUBTEST

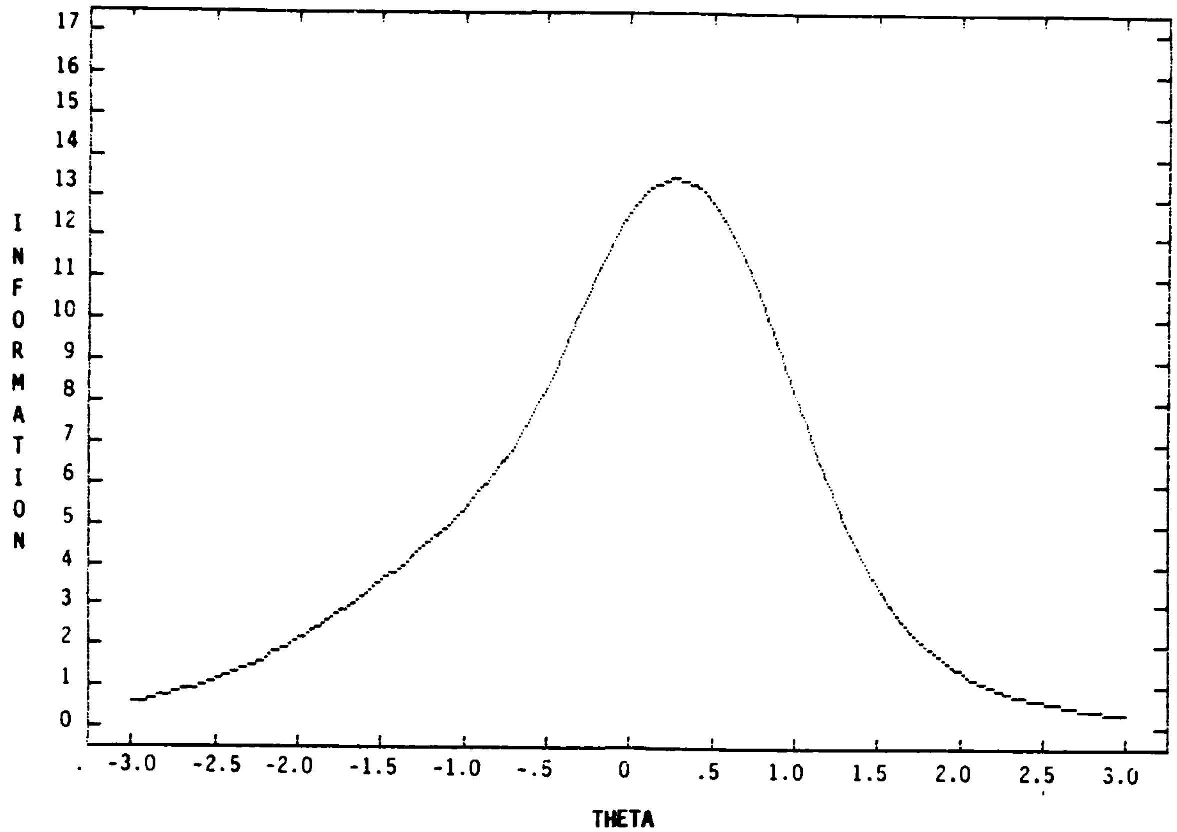


Figure C-13. TEST INFORMATION CURVE FOR READING COMPREHENSION SUBTEST

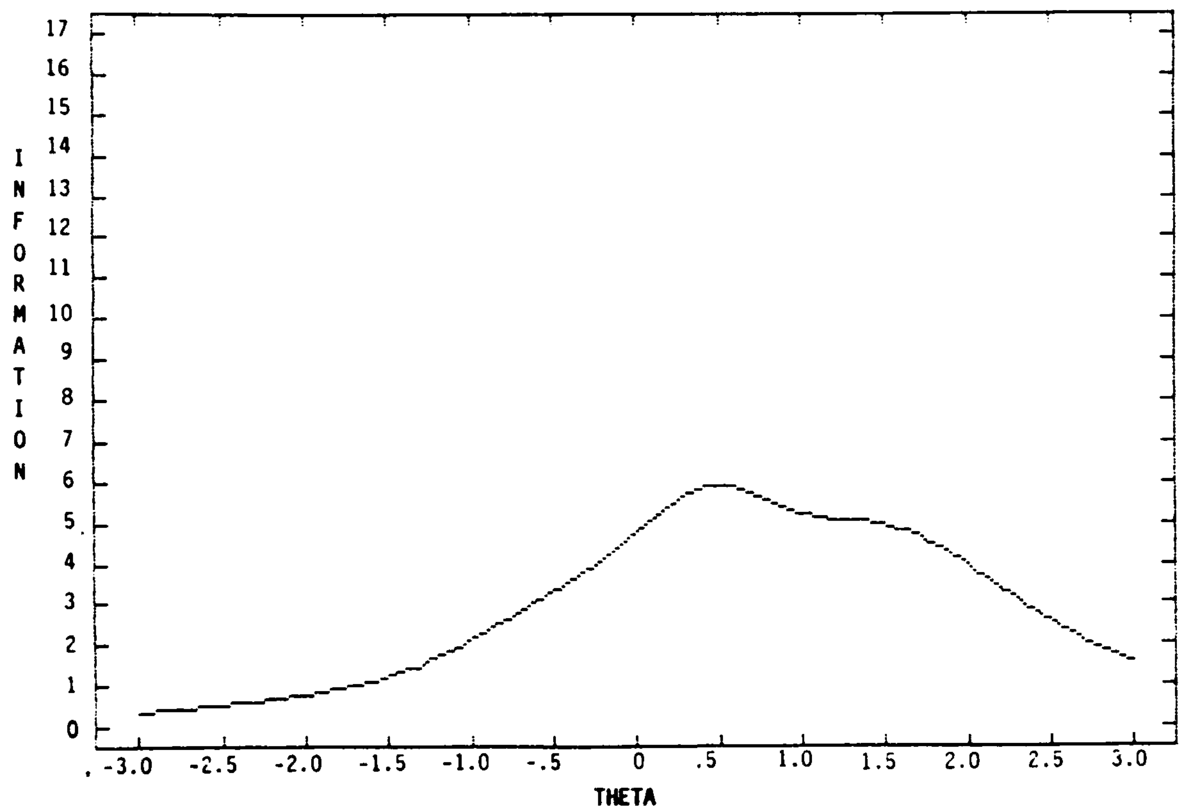


Figure C-14. TEST INFORMATION CURVE FOR DATA INTERPRETATION SUBTEST

APPENDIX D. DISTRIBUTIONS OF ABILITY OF THE SAMPLE ON THE SUBTESTS

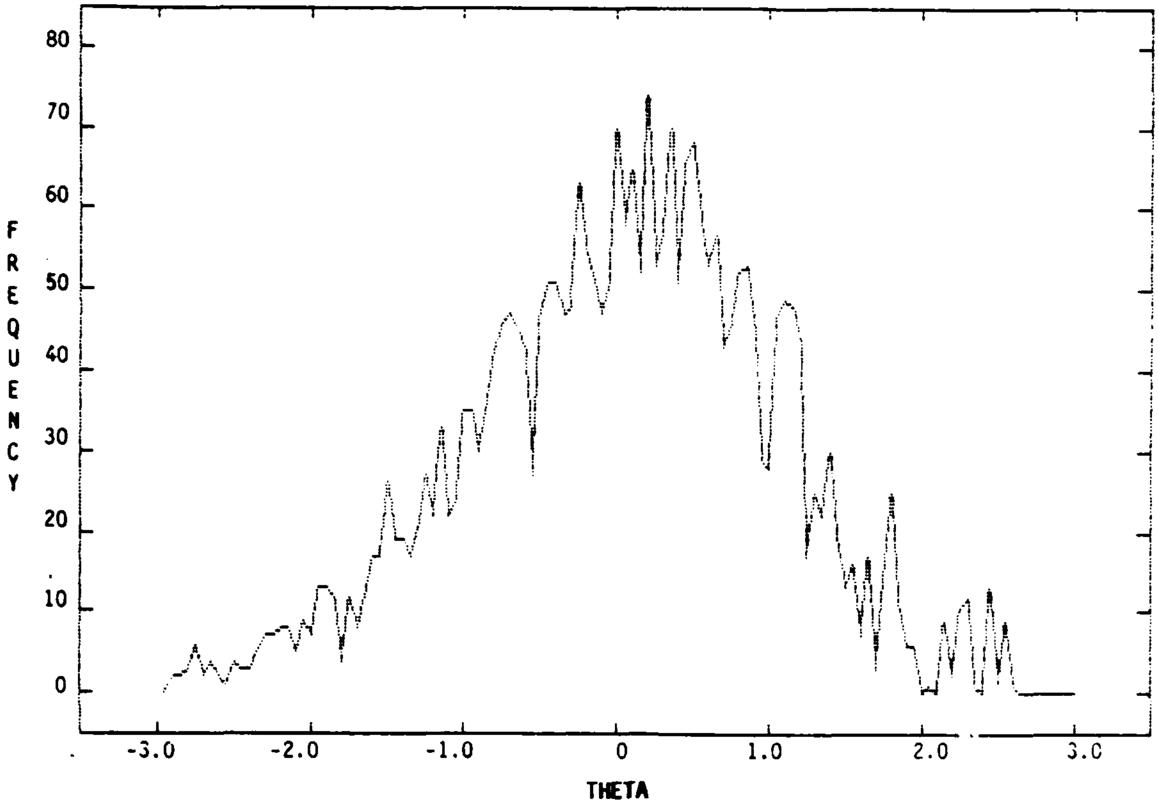


Figure D-1. DISTRIBUTION OF ABILITY (θ) FOR VERBAL ANALOGIES SUBTEST

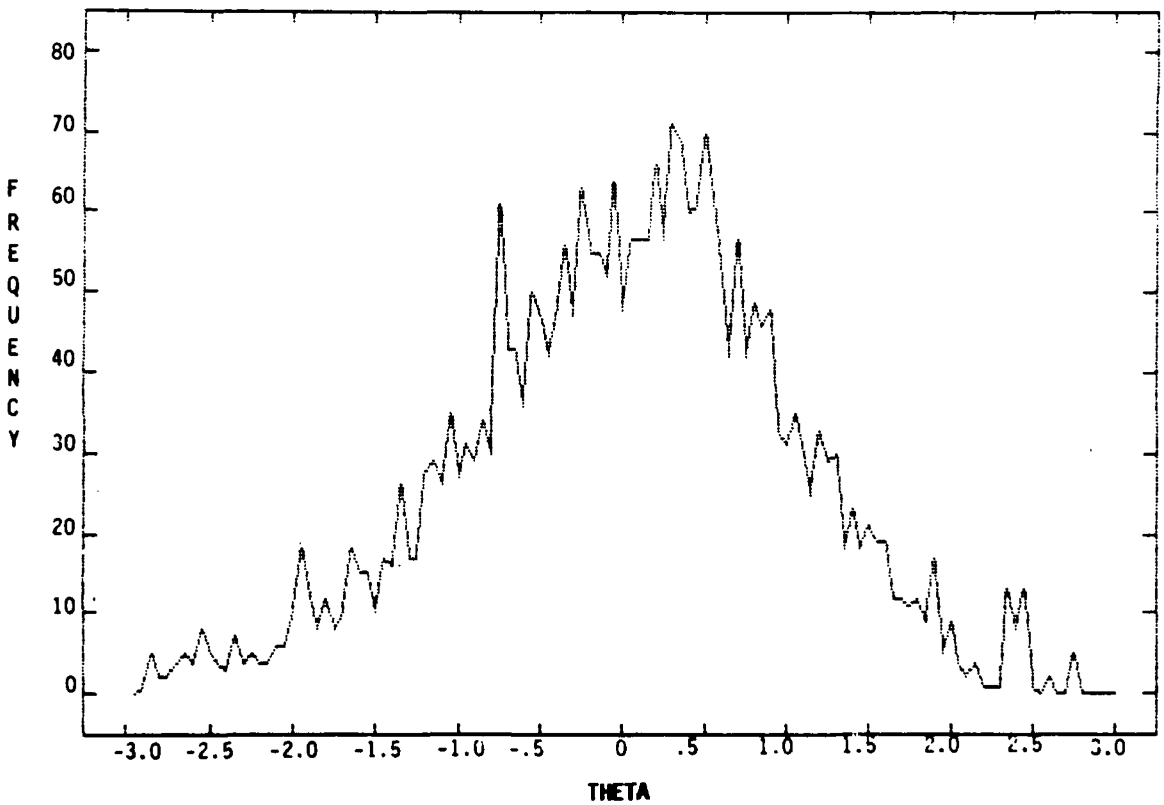


Figure D-2. DISTRIBUTION OF ABILITY (θ) FOR ARITHMETIC REASONING SUBTEST

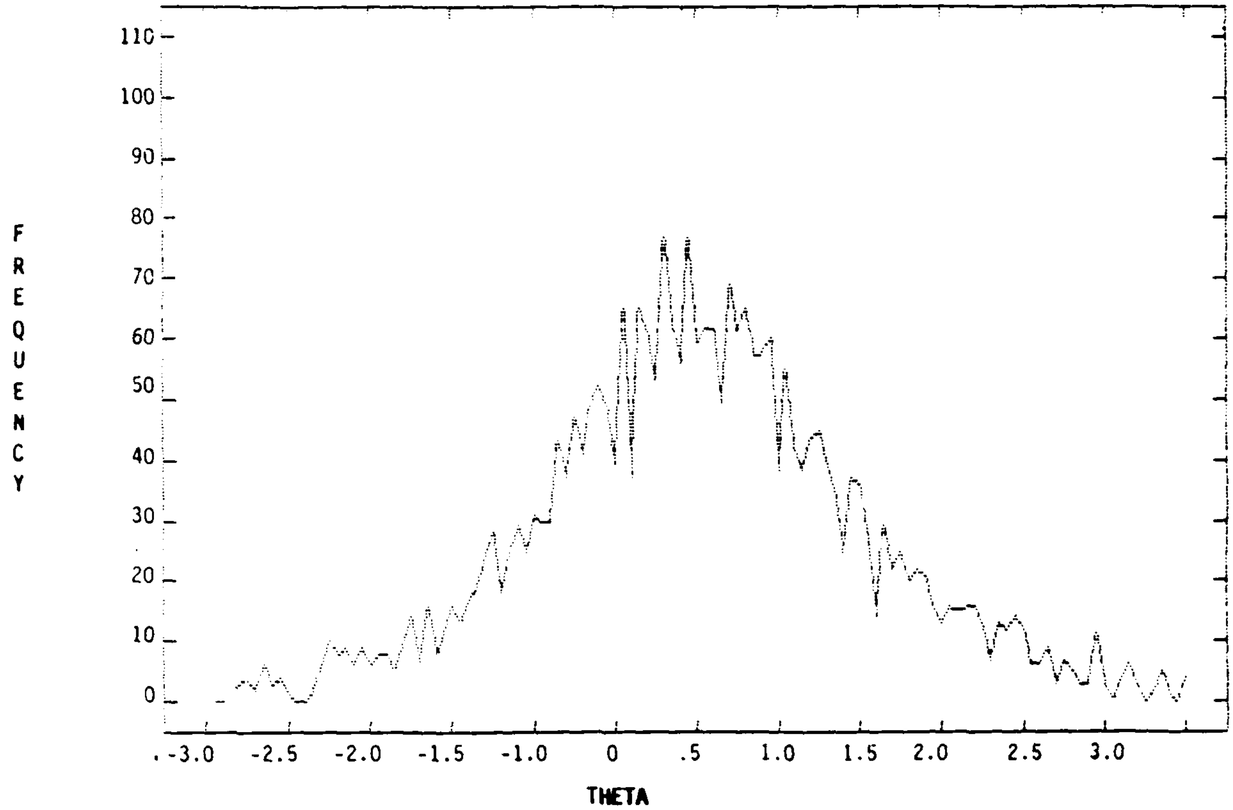


Figure D-3. DISTRIBUTION OF ABILITY (θ) FOR AVIATION INFORMATION SUBTEST

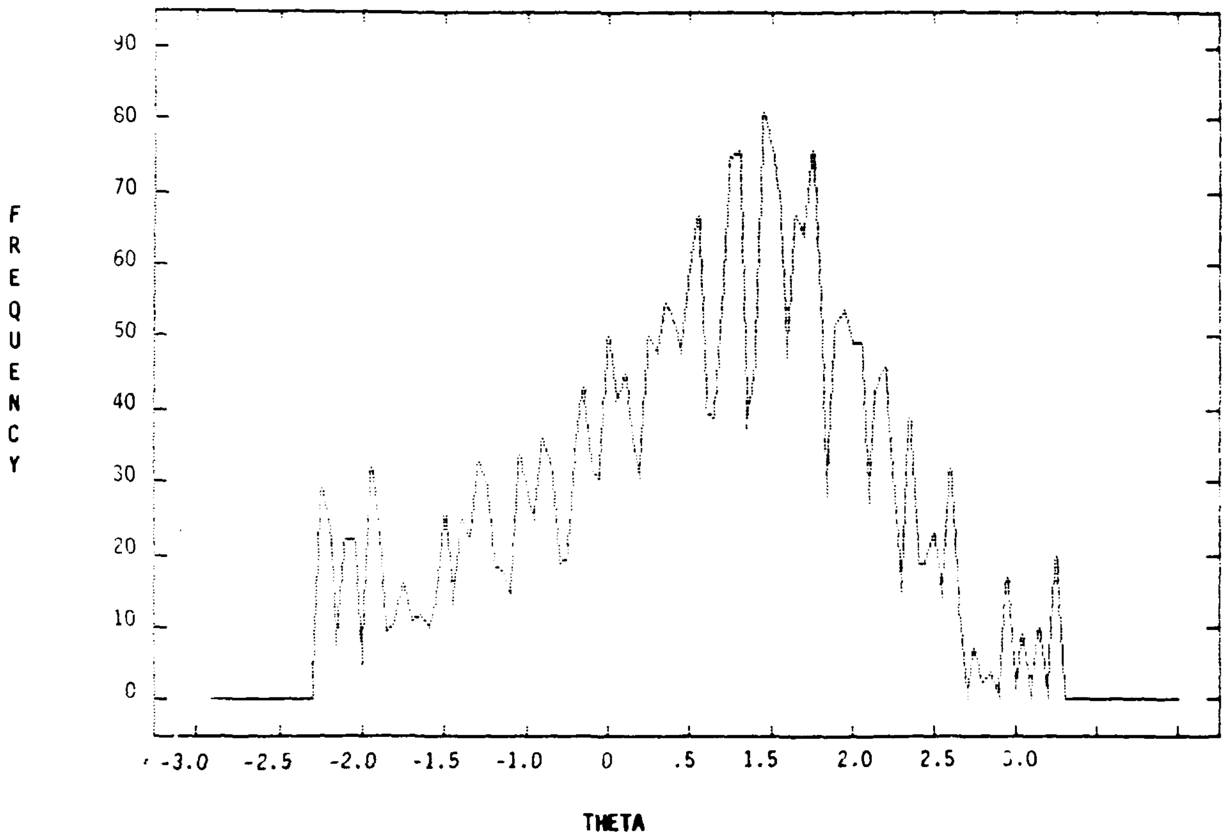


Figure D-4. DISTRIBUTION OF ABILITY (θ) FOR ROTATED BLOCKS SUBTEST

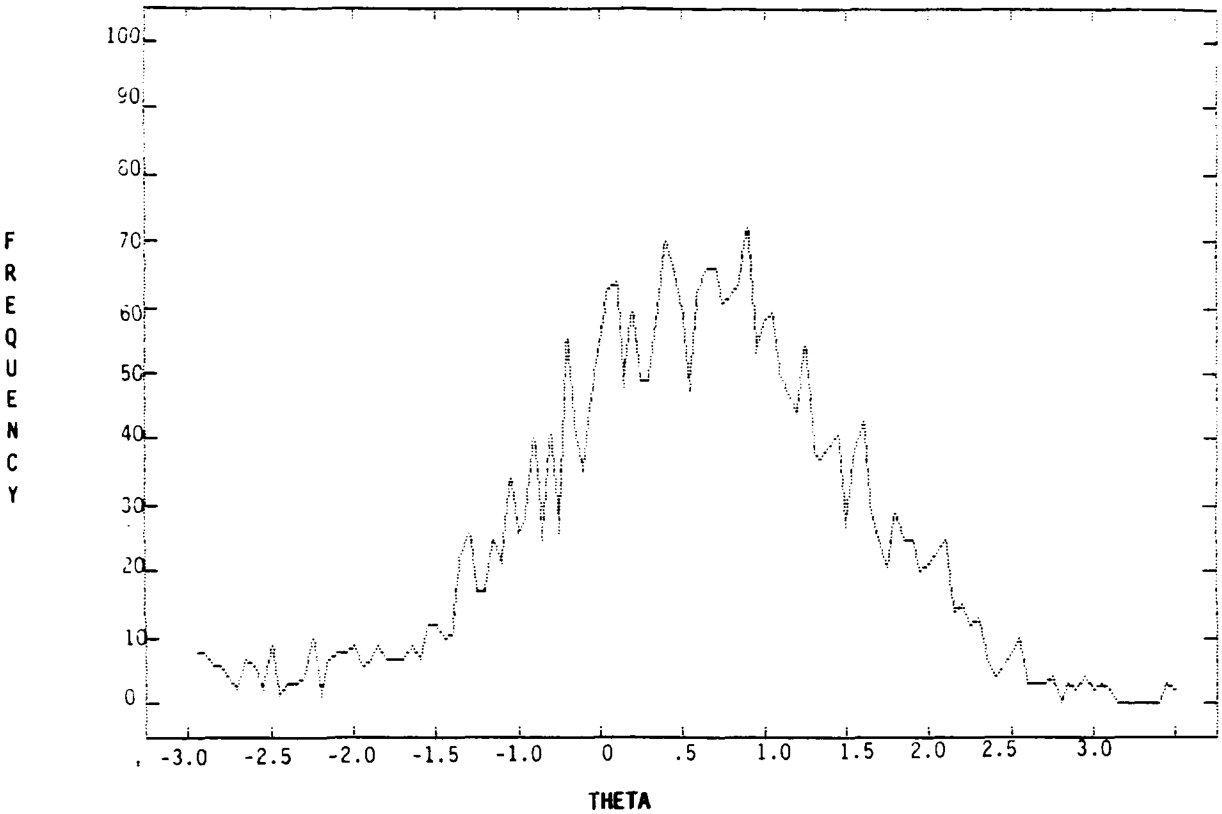


Figure D-5. DISTRIBUTION OF ABILITY (θ) FOR GENERAL SCIENCE SUBTEST

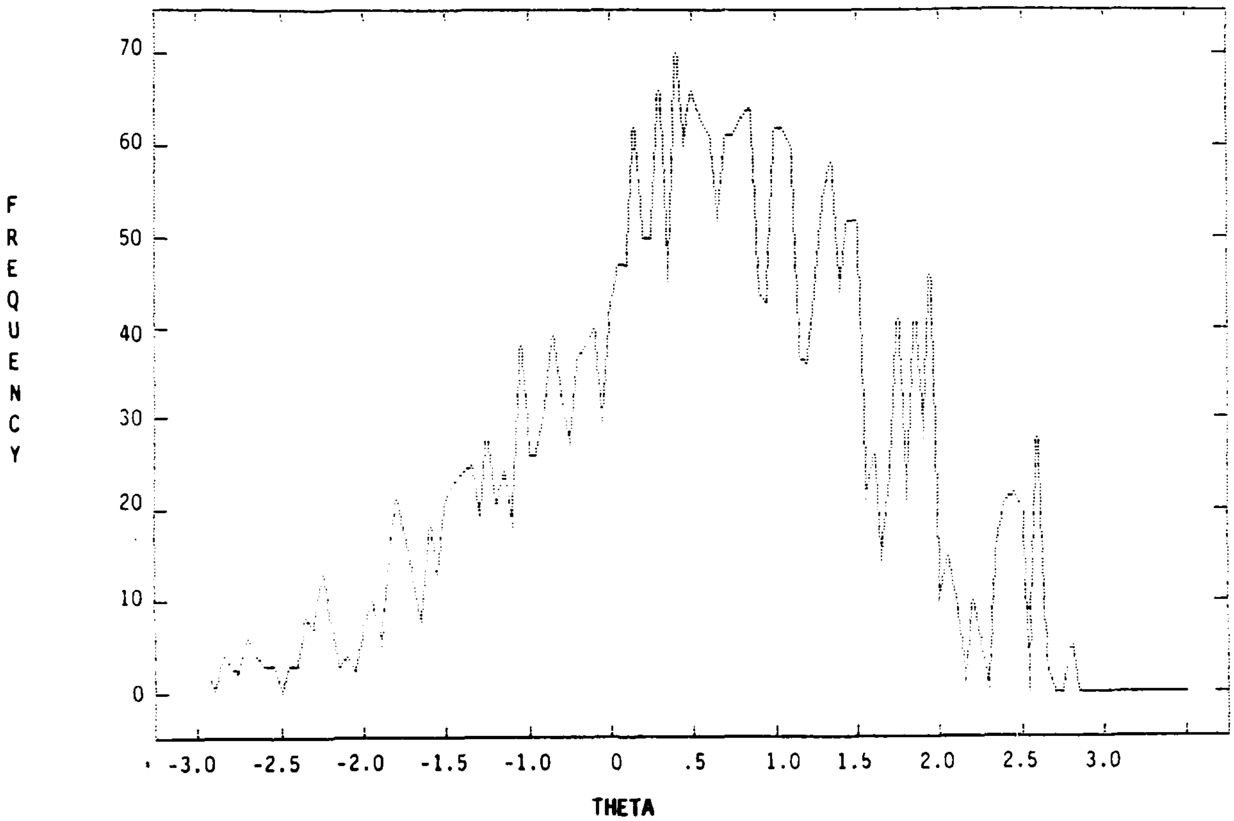


Figure D-6. DISTRIBUTION OF ABILITY (θ) FOR HIDDEN FIGURES SUBTEST

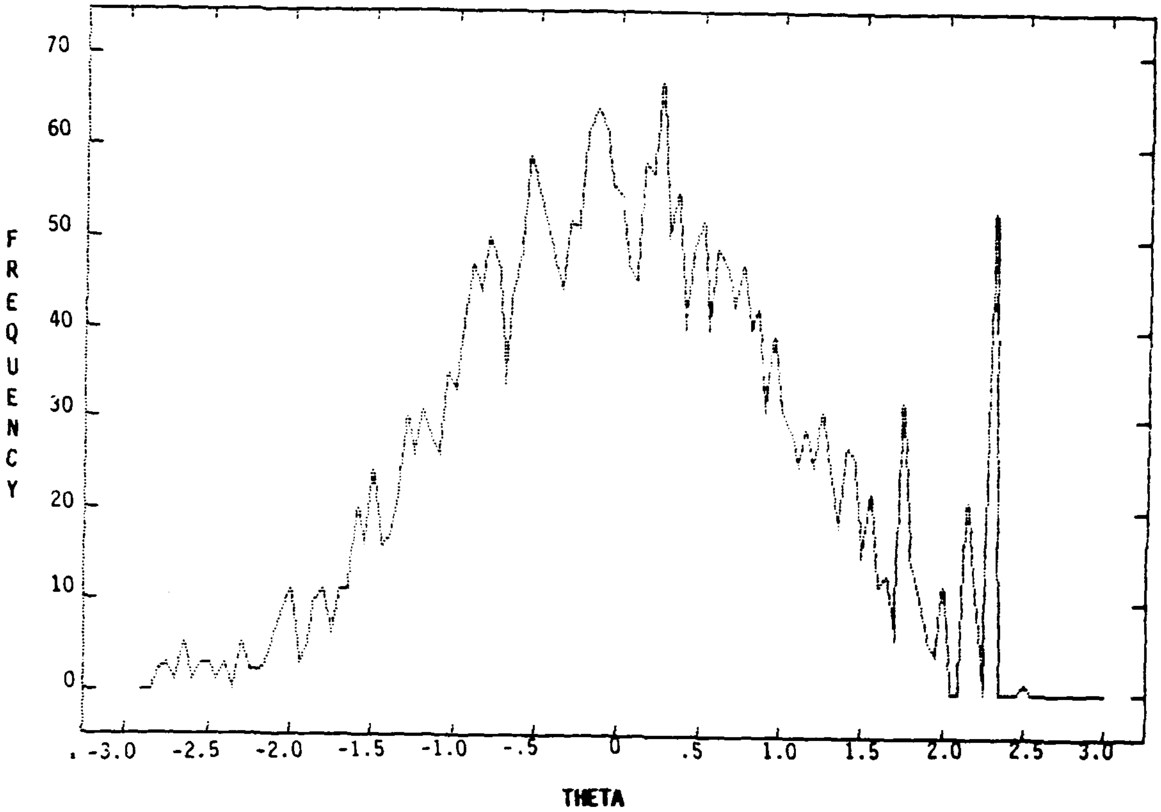


Figure D-7. DISTRIBUTION OF ABILITY (θ) FOR WORD KNOWLEDGE SUBTEST

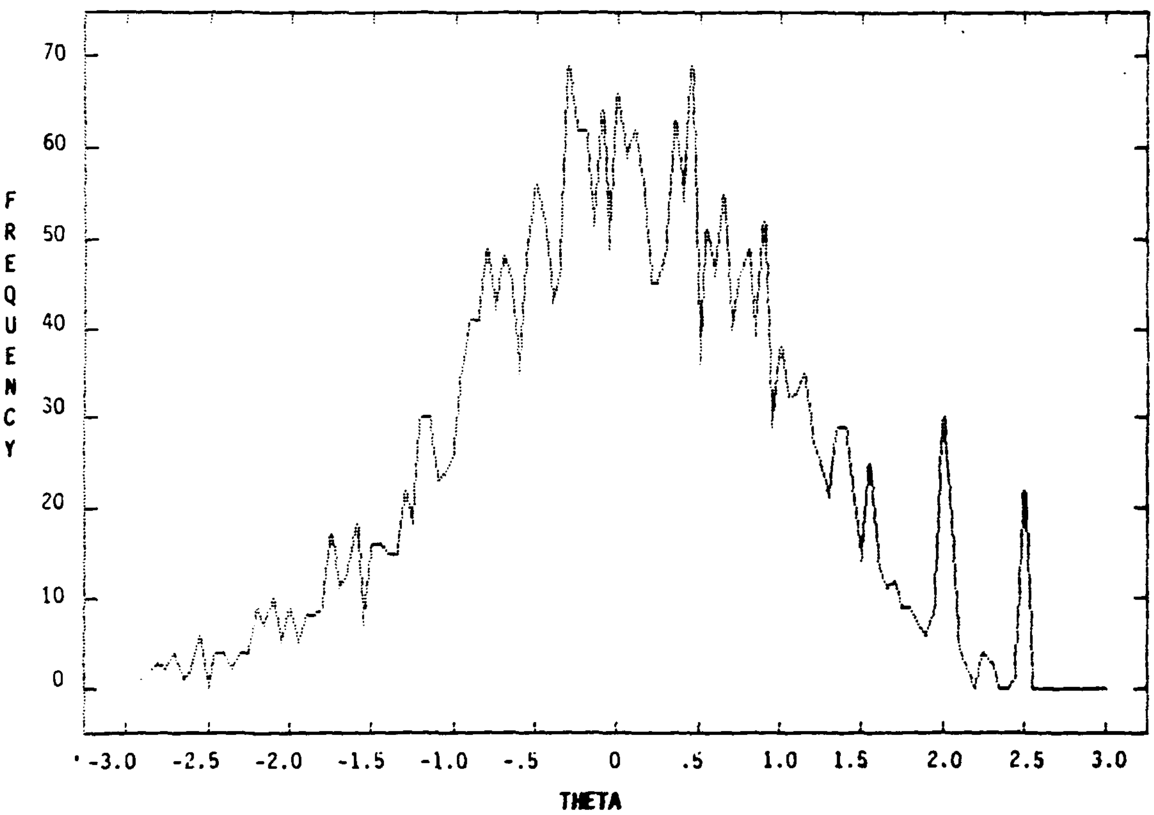


Figure D-8. DISTRIBUTION OF ABILITY (θ) FOR MATH KNOWLEDGE SUBTEST

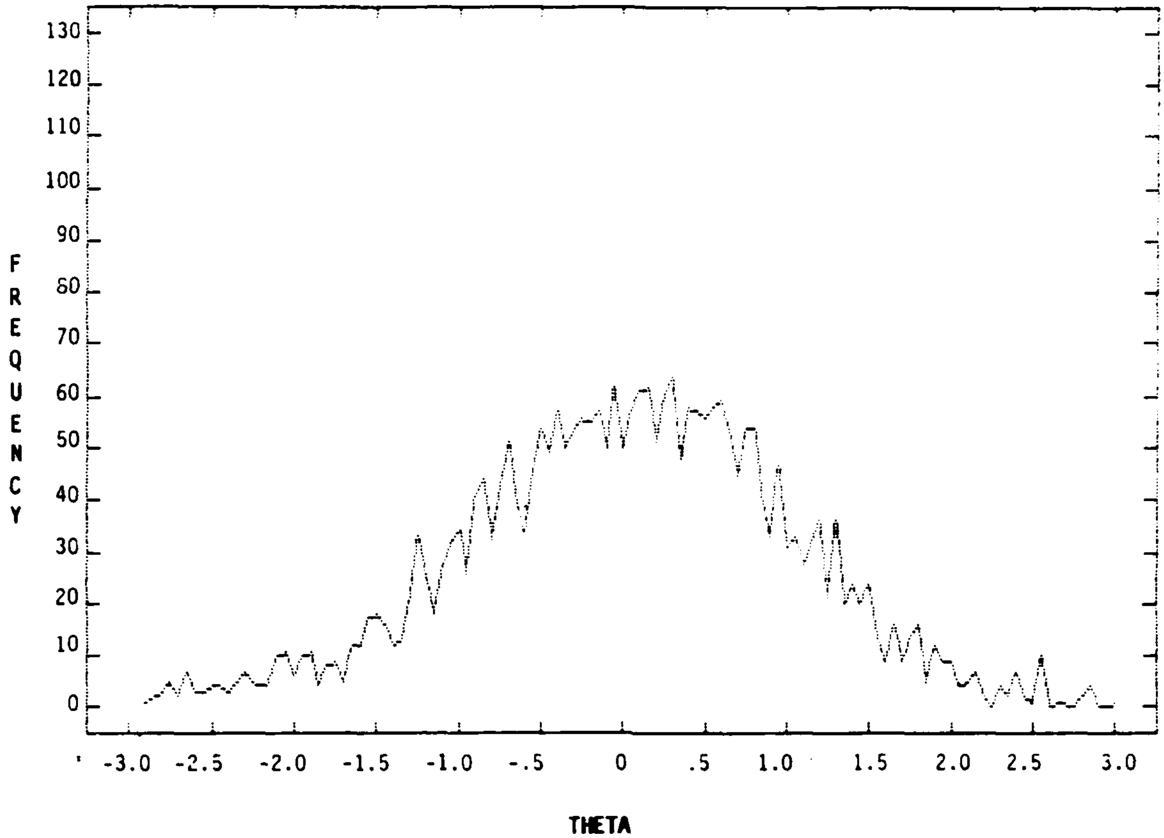


Figure D-9. DISTRIBUTION OF ABILITY (θ) FOR MECHANICAL COMPREHENSION SUBTEST

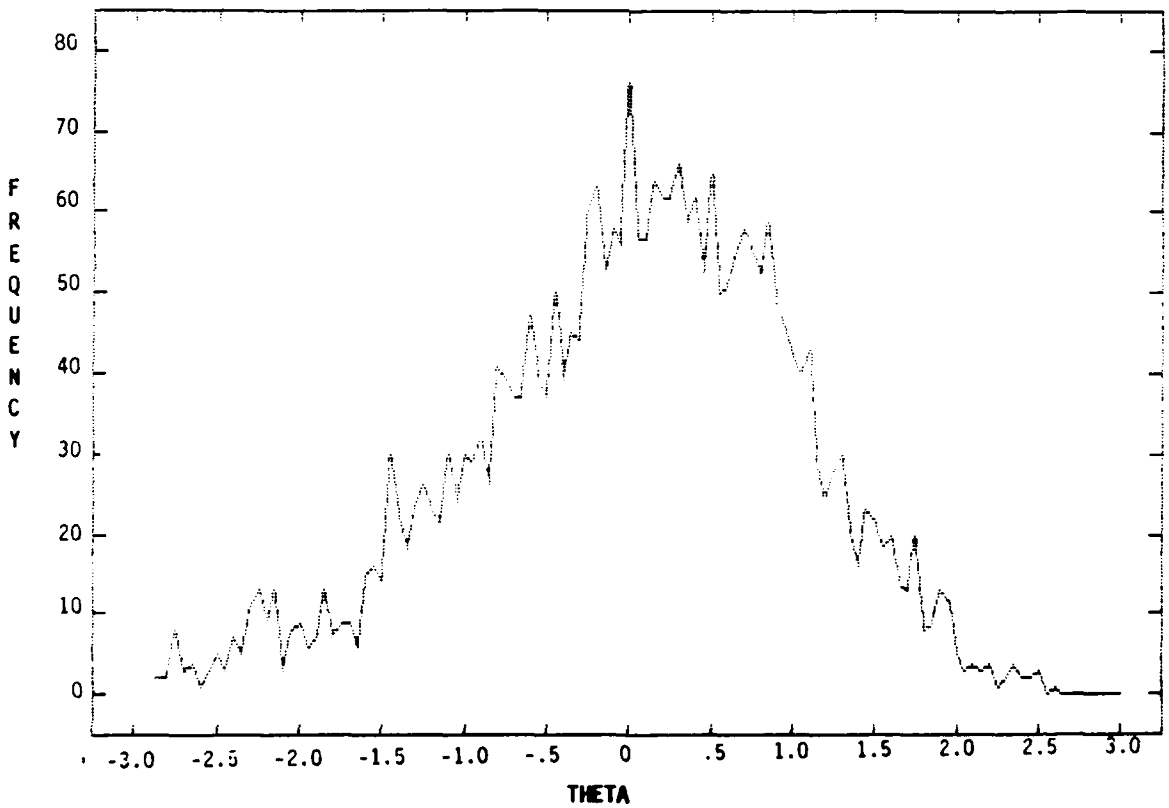


Figure D-10. DISTRIBUTION OF ABILITY (θ) FOR SCALE READING SUBTEST

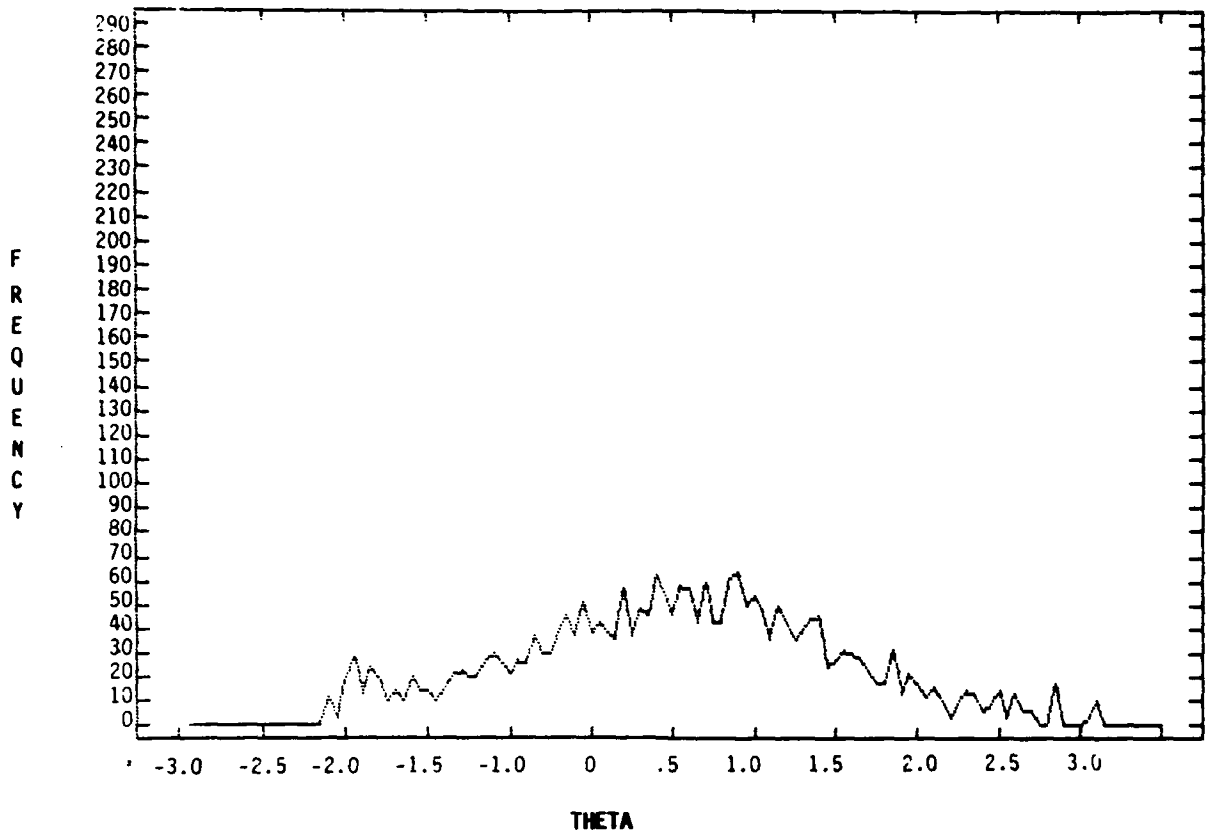


Figure D-11. DISTRIBUTION OF ABILITY (θ) FOR INSTRUMENT COMPREHENSION SUBTEST

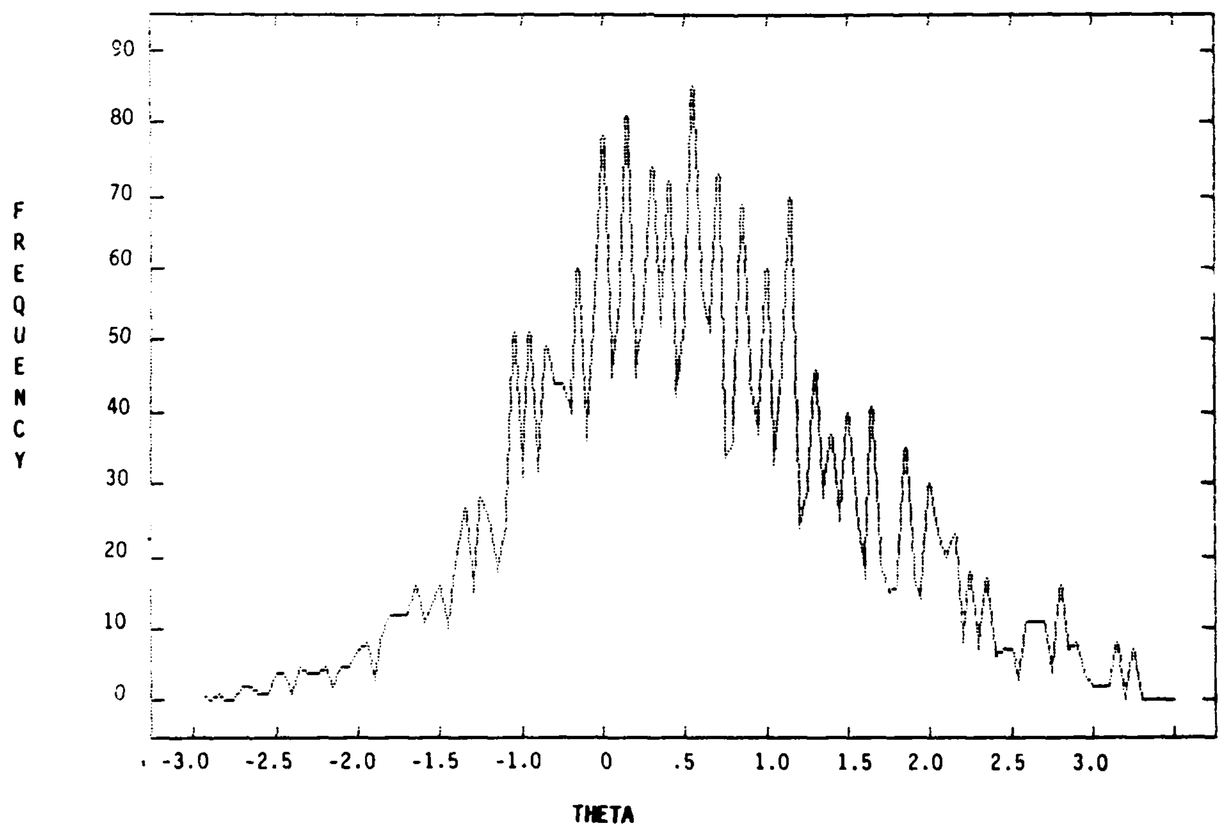


Figure D-12. DISTRIBUTION OF ABILITY (θ) FOR TABLE READING SUBTEST

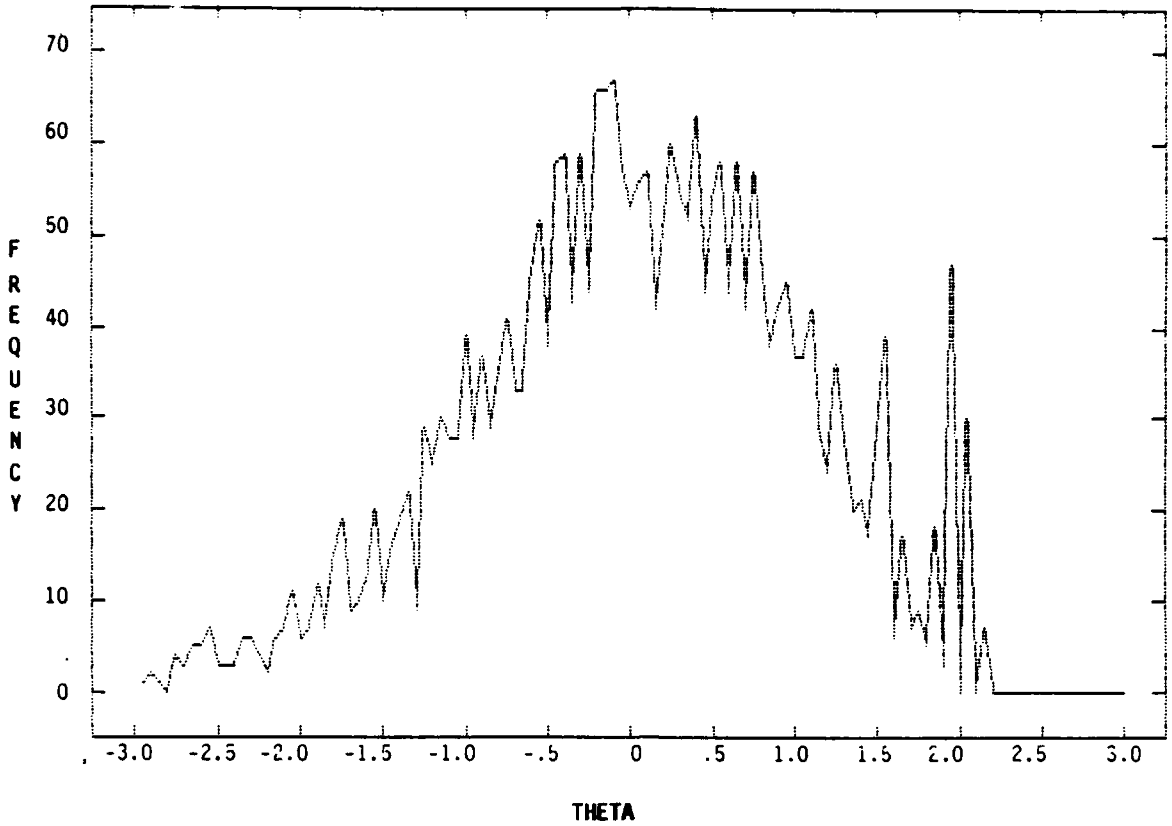


Figure D-13. DISTRIBUTION OF ABILITY (θ) FOR READING COMPREHENSION SUBTEST

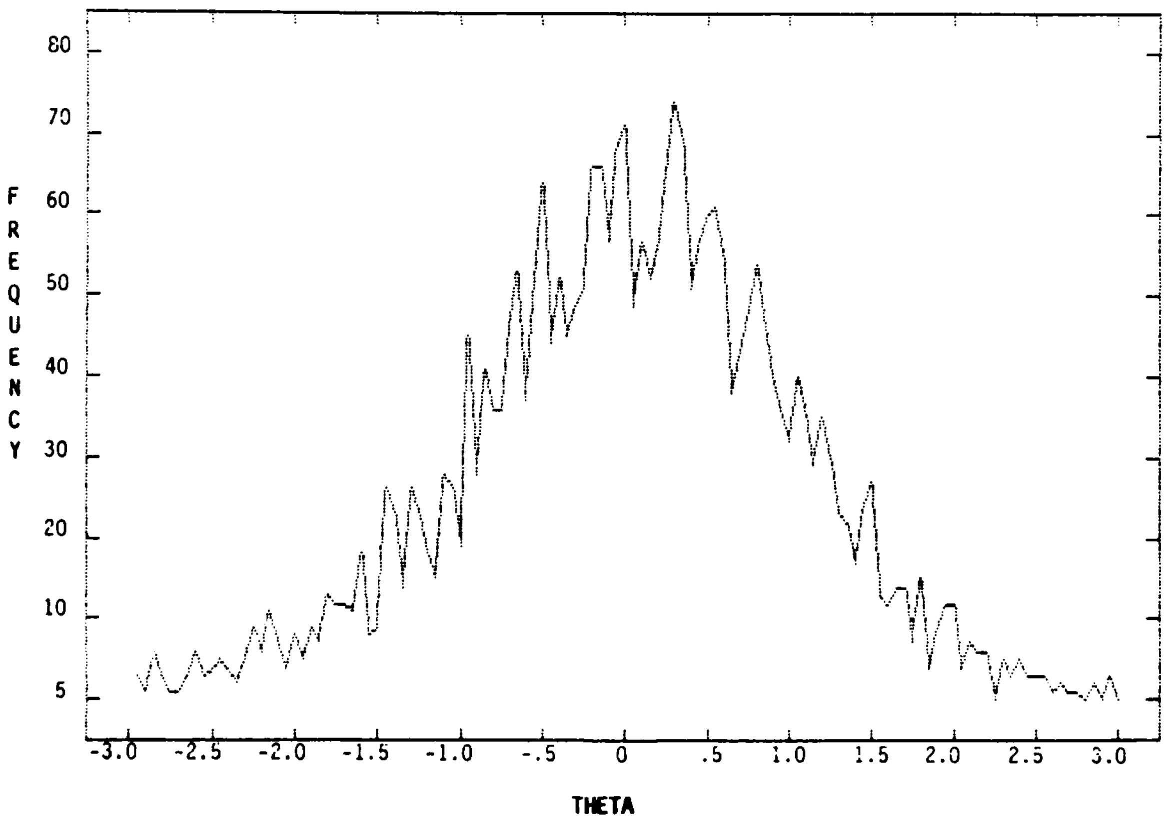


Figure D-14. DISTRIBUTION OF ABILITY (θ) FOR DATA INTERPRETATION SUBTEST

END
10-81
DTIC